Protein identification strategies for the greenshell mussel *Perna canaliculus*

By

Cassidy Moeke

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the requirements for the degree of
Master of Biomedical Science

Victoria University of Wellington
2012

Abstract

The greenshell mussel *Perna canaliculus* is considered to be a suitable biomonitor for heavy metal pollution. This is due to their ability to accumulate and tolerate heavy metals in their tissues. These characteristics make them useful for identifying protein biomarkers of heavy metal pollution, as well as proteins associated with heavy metal detoxification and homeostasis. However, the identification of such proteins is restricted by the greenshell mussel being poorly represented in sequence databases. Several strategies have previously been used to identify proteins in unsequenced species, but only one of these strategies has been applied to the greenshell mussel. The objective of this thesis was to examine different protein identification strategies using a combined two-dimensional gel electrophoresis and MALDI-TOF/TOF mass spectrometry approach.

The protein identification strategies used include a Mascot database search, as well as *de novo* sequencing approaches using PEAKS DB and SPIDER homology searches. In total, 155 protein spots were excised and a total of 68 identified. Fifty-six proteins were identified using a Mascot search against the Mollusca, NCBInr and Invertebrate EST database, with seven single-peptide identifications. *De novo* sequencing strategies identified additional proteins, with two from a PEAKS DB search and 10 from an error-tolerant SPIDER homology search. The most noticeable protein groups identified were cytoskeletal proteins, stress response proteins and those involved in protein biosynthesis. Actin and tubulin made up the bulk of the identifications, accounting for 39% of all proteins identified.

This multifaceted approach was shown to be useful for identifying proteins in the greenshell mussel *Perna canaliculus.* Mascot and PEAKS DB performed equally well, while the error-tolerant functionality of SPIDER was useful for identifying additional proteins. A subsequent search against the Invertebrate EST database was also found to be useful for identifying additional proteins. Despite this, more than half of all proteins remained unidentified. Most of these proteins either failed to produce good quality MS spectra or did not find a match to a sequence in the database. Future research should first focus on obtaining quality MS spectra for all proteins concerned and then examine other strategies that may be more suitable for identifying proteins for species with poor representation in sequence databases.

Acknowledgments

I would like to thank Dr. Bill Jordan for his continued guidance and support during this thesis project and Liz Richardson for the same reasons. I would also like to extend my thanks and gratitude to Dr. Paul Teesdale-Spittle for his input and guidance.

Special thanks go to Hannah Hoang and Sarah Cordiner for sharing their knowledge and expertise with the laboratory component of this project, and also to Dr. Jonathan Dunne for his all-round expertise with operating the mass spectrometer. I would also like to thank Danyl McLauchlan for his assistance during the bioinformatics phase of this experiment.

Final thanks go to the Foundation for Research, Science and Technology (now known as the Ministry of Science and Innovation) for funding this project and also to the research office of Victoria University of Wellington for their financial assistance.

Table of Contents

List of Figures

List of Tables

Abbreviations

2-DE	Two-dimensional gel electrophoresis
ACN	Acetonitrile
CHCA	α-cyano-4-hydroxycinnamic acid
CID	Collision induced dissociation
EST	Expressed sequence tag
IPG	Immobilized pH gradient
MALDI	Matrix-associated laser desorption/ionisation
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
*M*r	Molecular mass
m/z	Mass-to-charge ratio
PAGE	Polyacrylamide gel electrophoresis
PI	Isoelectric point
PMF	Peptide mass fingerprint
SDS	Sodium dodecyl sulfate

1. Introduction

1.1 *Perna canaliculus*: a biological monitor for heavy metal pollution

In the field of ecotoxicology, several studies have assessed the suitability of using mussels as a biomonitor for heavy metal pollution (Nicholson and Lam, 2005; Funes *et al.,* 2006; Al-Subiai *et al*., 2011; Baraj *et al.,* 2011). Although this research has in the most part centered on other mussel species, little research has involved the greenshell mussel *Perna canaliculus*. The characteristics which make mussels a suitable biomonitor for heavy metal pollution are their adaptive ability to accumulate and tolerate heavy metals in their tissues (Vosloo *et al.,* 2012). Not only do these characteristics make mussels useful for monitoring the health of the environment, but also for discovering new proteins associated with heavy metal tolerance and homeostasis. However, using the greenshell mussel *Perna canaliculus* for these purposes is restricted by limited ability to identify proteins in this species.

The only study that has investigated heavy metal bioaccumulation in *Perna canaliculus* demonstrates the difficulty with identifying proteins in the greenshell mussel. This study was carried out to identify protein biomarkers in response to treatment with mercury and cadmium (Whyte, 2006). In response to heavy metal treatment, 111 protein spots were selected as potential biomarkers. However, only two isoforms of tropomyosin and one actin isoform could be identified. The low number of successful protein identifications was attributed to the poor representation of mussels in sequence databases. All mussel species belonging to the Perna genus are still underrepresented in databases, containing a combined total of 14 protein sequences (www.ncbi.nlm.nih.gov). A workable strategy is therefore required to overcome these limitations.

Two commonly-used strategies for identifying proteins are the bottom-up and top-down proteomic approaches (Wehr, 2006). Bottom-up proteomics identifies proteins by first digesting proteins with a protease and then analysing the peptides using mass spectrometry (Yates *et al.,* 2009). A peptide mass fingerprint is generated, which can then be searched against a sequence database to identify the protein. Proteins can also be identified by acquiring amino acid related sequence information using tandem mass spectrometry. Top-down proteomics differs from bottom-up in that the intact protein is analysed (Chait, 2006).

Despite this, the bottom-up approach is considered to be the most preferred approach for identifying proteins.

1.2 Bottom-up proteomic approach

In mussels, bottom-up proteomics begins with a heterogeneous protein mixture from a tissue or organ sample (Tomanek and Zuzow, 2010; Leung *et al.,* 2011; Puerto *et al.,* 2011). These protein mixtures are usually complex, containing hundreds or thousands of proteins. Several different strategies are currently used for identifying proteins using bottom-up proteomics. One strategy involves analysing peptide mixtures using "shotgun" proteomics. This strategy requires the protein mixture to be first digested and the resulting peptides analysed by mass spectrometry (Marcotte, 2007). Proteins can then be identified by inference. Two drawbacks of using shotgun proteomics are the presence of isomeric peptides and high abundance peptides. Isomeric peptides are those that share a similar mass-to-charge ratio and cannot be told apart (Chen *et al*., 2010). High abundance peptides pose a problem since they can prevent low abundance peptides from being detected (Reiter *et al*., 2009). A strategy that can be used to minimise these issues is to first separate the peptide mixture prior to analysis.

Multi-dimensional protein identification technology is a strategy used for separating peptide mixtures. This strategy combines reverse-phase liquid chromatography with a strong cation exchange column to separate peptides based on their hydrophobicity and electrical charge (Delahunty and Yates III, 2007). Another strategy, called electrostatic repulsion hydrophilic interaction chromatography, can also be used to separate peptides. Electrostatic repulsion hydrophilic interaction chromatography is orthogonal to reverse-phase liquid chromatography and uses electrostatic repulsion and hydrophilic interactions to separate peptides according to their isoelectric point and polarity (Hao *et al.,* 2011). Separating peptide mixtures improves the scope for inferring protein identities, but degenerate peptides and "one-hit wonders" can cause protein identifications to be ambiguous (Nesvizhskii and Aebersold, 2005). Instead, a more appropriate approach may be to first separate the protein mixture prior to analysis.

Two-dimensional gel electrophoresis (2-DE) is a commonly used technology for separating protein mixtures. It is well known for its powerful protein separation capabilities and high resolving capacity (Penque, 2009). 2-DE first separates proteins according to their isoelectric point and then by their molecular mass. Not only does 2-DE provide additional information to assist with protein identifications, but it is also effective for identifying proteins by peptide mass fingerprinting (Rabilloud and Lelong, 2011). Protein identification strategies using 2-DE have previously been applied to several mussel species, with each study demonstrating mixed successes (discussed further in section 1.4). A common 2-DE proteomics workflow typically involves first separating a heterogeneous protein mixture. Proteins are then selectively isolated, digested and analysed by mass spectrometry.

1.2.1 2-DE separation of a heterogeneous protein mixture

2-DE separates proteins in two dimensions. The first dimension involves separating proteins according to their isoelectric point (pI). Proteins carry a charge that can be influenced by the pH of the surrounding medium. Upon applying an electrical current, proteins can migrate through a pH gradient until their net charge becomes zero. The specific pH where the protein becomes stationary is known as their pI (Cargile *et al*., 2004). Proteins with a high pI value have a greater amount of basic amino acid residues than those with a lower pI value (Kiraga *et al.,* 2007). Commercially available immobilized pH gradient (IPG) strips can be used to help separate proteins in the first dimension. IPG strips contain a gradient of acidic and basic buffer groups fixed to a polyacrylamide gel (Vercauteren *et al.,* 2007). By fixing these groups into place helps overcome reproducibility issues associated with carrier-ampholyte pH gradients.

The second dimension involves separating proteins based on their molecular mass (*M*r). Before proteins enter the second dimension, they should first be denatured. Sodium dodecyl sulphate is a commonly used protein denaturant that provides each protein with a net negative charge that is proportional to its mass (Clark, 2009). This allows proteins to be separated using a pore-based gel system, where smaller proteins migrate faster than larger proteins under the influence of an electric field (Garfin, 2003). Resolved proteins can then be visualised by staining with Coomassie brilliant blue, which allows proteins to be

selectively isolated. 2-DE can be used to resolve hundreds to thousands of proteins (Weiss and Görg, 2009).

1.2.2 In-gel protease digestion of proteins separated by PAGE

In-gel digestion uses a protease to hydrolyse at peptide bonds in proteins. One commonly used protease is trypsin. Trypsin diffuses into the gel and hydrolyses proteins at the carboxyl side of lysine and arginine residues, unless they are followed by proline. This creates a set of peptides that can then be analysed by mass spectrometry. Trypsin cleaves proteins with high specificity, which minimizes the occurrence of non specific cleavage products. This is achieved in the binding site of trypsin by a negatively charged aspartate residue being accessible only by positively charged amino acids with long side chains (Olsen *et al*., 2004). In addition to being highly specific, trypsin typically produces peptides in the preferred mass range of 800 to 4000 Da for mass spectrometry analysis (Tran *et al*., 2011).

1.2.3 MALDI-TOF/TOF mass spectrometry

Mass spectrometry (MS) is used to measure the mass-to-charge (m/z) ratio of ionised peptides. Matrix-assisted laser desorption ionisation (MALDI) is a commonly used soft ionisation source that assists the peptide into the gas phase (Domon and Aebersold, 2006). The matrix absorbs UV light from a laser source, leading to gas phase matrix and peptide ions (Karas *et al.,* 2000). A suitable matrix for MALDI is *alpha*-cyano-4-hydroxycinnamic acid (CHCA), which is regarded as the gold standard for peptide analysis (Beavis *et al.,* 1992). CHCA has a reduced tendency to discriminate against peptides of different amino acid compositions. This is partly due to CHCA favouring peptides with arginine residues, which are one of the two dominant peptide types produced during a trypsin digest.

Mass spectrometry requires a mass analyser to measure the m/z ratio of ionised peptides. A common mass analyser used in combination with MALDI is time-of-flight (TOF). MALDI-TOF MS measures the velocity of ionised peptides to determine their m/z ratio. Ionised peptides travel though a flight tube where they are separated according to their flight time, with

smaller ions reaching the detector reflects than larger ions. The time it takes for each peptide to reach the detector determines their mass, while peptides analysed using the normal conditions in the MALDI procedure each carry a single charge (Aebersold and Mann, 2003; Cotter *et al.,* 2005). The m/z ratio of each peptide is then used to generate a peptide mass fingerprint, which can then be searched against a database to identify the protein (discussed further in section 1.3). In some cases, the peptide mass fingerprint can be insufficient for successful protein identification. In these instances, tandem mass spectrometry can be used.

Tandem mass spectrometry (MS/MS) is used to obtain sequence-related information for peptides. MS/MS can be achieved using MALDI instruments coupled to a dual time of flight mass analyzer (TOF/TOF). MALDI-TOF/TOF MS/MS acquires sequence related information by fragmenting peptides using collision-induced dissociation (CID) (Papayannopoulos, 1995). CID can either be carried out using low or high energy, with low energy CID cleaving at the peptide backbone of the C- or N-terminus. C-terminal cleavage creates predominantly so-called y-series ions, while N-terminal cleavage creates ions of the so-called b-series. Whether y- or b-series ions are favoured depends on the amino acid composition of each peptide (Khatun *et al*., 2007). High energy CID also generates y- and b-series ions, but additionally cleaves the amino acid side-chain to create non-specific cleavage products (Griffiths *et al*., 2001).

1.3 Strategies for identifying proteins

Several different strategies are currently used to identify proteins. Peptide mass fingerprinting (PMF) is a common strategy used to identify proteins separated by 2-DE, in-gel digested and analysed by MALDI-TOF/TOF mass spectrometry. PMF involves searching a peaks list containing a set of peptide masses against a theoretical mass list constructed using a protein sequence database (Thiede *et al*., 2005). A theoretical mass list is generated by performing a theoretical digest on protein sequences using specified criteria that closely resembles the experimental conditions. Criteria include the enzyme used, allowed number of missed cleavages, mass error tolerance and common modifications. Missed cleavages arise from an incomplete enzymatic digest, while modifications can either result from *ex vivo* or *in vivo* events (Webster and Oxley, 2005). PMF is useful for quickly identifying proteins purified by 2-DE, but only if the annotated protein sequence is available in databases (Pappin *et al.,* 1993). If a species is poorly characterized in sequence databases, then cross-species identification may provide clues to identity.

Cross-species identification makes use of annotated sequences from other species to identify proteins. For cross-species identification to be effective, a high degree of amino acid sequence identity needs to be shared between proteins (Wilkins and Williams, 1997). This is often the case for highly conserved proteins or for those belonging to a closely-related species (Liska and Shevchenko, 2003). A major problem with cross-species identification is when protein sequences are too dissimilar for an accurate comparison to be made. Even a single amino acid difference is enough to prevent proteins from being identified by PMF (Wright *et al*., 2010). Protein isoforms, spliced variants and mass shifts induced by modifications and disulphide bridges, can also affect protein identifications (Thiede *et al.,* 2005). To help overcome these issues, MS/MS identification strategies can be used.

Proteins identified using MS/MS spectra usually fall into one of two categories. The first involves searching peptide masses against a protein database to select protein candidates for scoring. Fragment ions are then assigned to each candidate to identify the protein (Edwards, 2011). This combined approach has been shown to be more effective than using PMF alone (Wasinger *et al*., 1995; Wilkins and Williams, 1997). For species inadequately represented in protein databases, this search can also include an expressed sequence tag

(EST) database. EST sequences are short, single-pass copies of messenger RNA which are used to encode proteins. Although EST sequences are prone to errors, they can be useful as a supplementary search option for identifying novel proteins (Edwards, 2007). A second strategy involves using MS/MS spectra to construct *de novo* sequence tags. *De novo* sequence tags can then be searched against a protein database or used in a homology search (Ma and Johnson, 2012). This hybrid approach is regarded as the only alternative for identifying proteins that cannot be identified by a database search (discussed further below).

Validating protein identifications is an important task for detecting false-positive identifications. False-positive identifications are incorrect protein assignments that can arise from contaminating peptide and chemical sources, or isomeric peptides. Prior knowledge of contaminant masses, a data refinement step or a search against a contaminants database can help to reduce false-positive identifications (Sadygov *et al.,* 2004). The standard procedure for validating identifications is a target-decoy search. This involves re-searching the protein database, but with its sequences reversed or randomised (Elias and Gygi, 2007; Barboza *et al.,* 2011). However, it is believed an incorrect estimation of the false-positive rate is given when a small dataset is used. Instead, protein identifications can be validated using an independent *de novo*-based search strategy (Rogers *et al*., 2004).

1.3.1 Database searching using Mascot

Mascot is a widely-used database search engine that implements a probability-based scoring algorithm. This search engine identifies proteins by searching both peptide masses and MS/MS spectra against a database. Previous studies have shown as little as three peptide masses are needed to identify proteins by PMF, while only a single peptide is required for a MS/MS search (Laukens *et al*., 2004; Pappin *et al.,* 1993). In Mascot, a peaks list containing peptide masses are submitted to either identify proteins by PMF or for selecting a pool of protein candidates for scoring. MS/MS spectra are then searched against these protein candidates to compute an identity score which is dependent on the number and quality of fragment ion assignments (Perkins *et al.,* 1999). Finally, identity scores are

compared to a significance threshold to determine the likelihood the match arises from a chance event.

1.3.2 PEAKS Studio 5.3

PEAKS Studio 5.3 is a proteomic software package with three core functionalities: PEAKS *de novo* sequencing, PEAKS DB and SPIDER homology search. PEAKS *de novo* sequencing is an automated approach that uses MS/MS spectra to construct *de novo* sequences independently of databases. These *de novo* sequences are then used to infer protein identities using PEAKS DB or SPIDER. *De novo* sequencing begins with a pre-processing step to ensure only good quality spectra are retained (Ma *et al.,* 2003). Thousands of sequences are then constructed, with the highest scoring matches selected for confidence scoring (Zhang *et al,* 2011). Confidence scores are applied to each sequence tag, as well as to the positional confidence of each residue. Despite this, *de novo* sequences still remain prone to errors due to amino acids that share a similar mass.

PEAKS DB is used for searching *de novo* sequence tags against a database. *De novo* sequence tags are first searched against a protein database to create a pool of protein candidates (Zhang *et al.,* 2011). This helps reduce the size of the searchable database and allows MS/MS spectra to be efficiently searched against each candidate to select peptides for scoring. Peptides are scored according to several factors, including shared similarities with the *de novo* sequence, peptide length and error tolerance, with the highest-scoring peptides used to infer protein identifies. PEAKS has previously been successfully applied when a conventional Mascot database search was unsuccessful (Tannu and Hemby, 2007). In this study, 13 out of 30 proteins were unambiguously identified.

SPIDER is another option for identifying proteins using *de novo* sequences. This search strategy provides an error-tolerant homology search option for proteins unidentified after a PEAKS DB search. SPIDER differs from PEAKS DB in that it does not penalise mismatches arising from *de novo* sequence errors. *De novo* errors come about due to amino acids that share a similar mass, such as leucine and isoleucine or lysine and glutamine. SPIDER compensates for this by regarding these amino acids to be identical to each another.

Another feature of SPIDER is it allows for insertion, substitution and deletion mutations during a search. (Yuen, 2011). By allowing for mutations, this search option may prove useful for identifying proteins in a species with little sequence content in databases.

1.4 Protein identification studies for poorly characterised species

Protein identification is an integral part of proteomics. Not only does it help identify potential protein biomarkers, but also novel protein candidates (Brosch *et al.,* 2011). Since protein identification relies heavily on the amount of comparable sequence content, successful proteomic experiments are usually limited to species that are adequately represented in sequence databases (Wright *et al.,* 2010). This limitation can be detrimental for researching responses to specific stimuli in species such as the greenshell mussel *Perna canaliculus*. Several proteomic studies have previously been carried out in species with limited protein content in databases, all demonstrating mixed results. This section will consider the different strategies used in these studies, as well as the successes and problems encountered.

1.4.1 Identifying proteins from mussels

One of the earliest protein identification studies in mussels was carried out to create a protein reference map for *Mytilus edulis* and *Mytilus galloprovincialis*. This study used foot tissue and selected 37 differentially expressed proteins to be identified. Fourteen of these proteins were identified by PMF, while only a single protein was identified by MS/MS (López *et al*., 2002). Failure to identify the remaining 22 proteins was considered to be the result of the low number of comparable sequences in databases. A second study carried out a few years later also encountered similar issues. In this study, 132 proteins were selected for identification (Manduzio *et al.,* 2005). Despite acquiring good MS and MS/MS spectra, only 19 were identified. Nearly all of these proteins were conserved and identified by cross-species identification.

More recent studies had greater success. One study investigated the effects of temperature on the gill proteome for *Mytilus galloprovincialis* and *Mytilus trossulus* (Tomanek and Zuzow, 2010). This study used 2-DE to resolve proteins which were digesting using trypsin and analysed by MALDI-TOF/TOF mass spectrometry. MS and MS/MS spectra were combined in a Mascot database search against a Mollusca protein sequence database and a Mytilus-specific EST database. This strategy was considerably more successful than previous strategies, with a total of 108 proteins being identified. A second study demonstrated improved success using an NCBInr database search to identify 41 proteins (Letendre *et al.,* 2011). Not only can these improved outcomes be attributed to the increase in size of protein sequence databases, but also the incorporation of an EST database search.

Although most protein identification studies have involved species of the Mytilus genus, one recently published study was carried out using the mussel *Perna viridis*. This study was carried out to assess the impact of cadmium and hydrogen peroxide on the proteome of the hepatopancreas and adductor muscles. Using a 2-DE approach, 37 proteins were selected for analysis by MALDI-TOF/TOF mass spectrometry. A database search was carried out using the MS/MS ions functionality of Mascot, where each protein would be searched against both an "other metazoan" and Invertebrate EST database. Of these proteins, 15 were identified with more than half arising from the Invertebrate EST database search (Leung, Wang *et al*., 2011). Protein identification studies involving other bivalve species have also shown limited success, with six proteins identified for *Corbicula fluminea* and seven for *Dreissena polymorpha* (De Souza *et al.,* 2009; Puerto *et al*., 2011).

1.4.2 Identifying proteins in non-bivalve species

Protein identification studies involving other non-sequenced species have also demonstrated mixed results. In these studies, a *de novo* sequencing approach was used followed by a homology search. A study carried out in the bacterium *Halorhodospira halophila* managed to identify 31 proteins using this approach (Samyn *et al.,* 2006), while another study involving bell pepper, spinach and cassava identified 45, 44 and 31 proteins, respectively (Grossmann *et al*., 2007). However, not all studies have shared these successes.

One of these studies could only identify six proteins in the green algae *Dunaliella salina* (Waridel *et al.,* 2007), while a separate study failed to identify even a single protein (Martínez-Fernández *et al*., 2008). Low protein abundance and insufficient protein content in databases were given as the reason for the low number of proteins identified.

Protein sequence content plays a major role for identifying proteins. One way to understand the significance of having sufficient sequence content is to compare studies using a species with a recently sequenced genome. One such species is the bacterium *Pseudomonas putida.* Prior to its genome being sequence, only three out of 100 randomly selected proteins could be identified (Krayl *et al*., 2003; Monsinjon and Knigge, 2007). However, once the genome was sequenced 195 proteins were identified in a single experiment. Another study demonstrated the major benefits that can be gained when using a sequenced genome from a closely-related species. In this study, the fully sequenced genome of *Daphnia pulex* was used to identify proteins from *Daphnia longicephala*. This search resulted in the successful identification of 371 proteins (Fröhlich *et al*., 2009). But when the complete sequence database of *Drosophila melanogaster* was used, only 71 proteins could be identified. This shows how useful sequences from a closely-related species can be for identifying proteins in poorly characterised species.

1.5 Research objectives

The goals of this thesis were to examine different strategies for identifying proteins in the greenshell mussel *Perna canaliculus.* Protein spots were excised from 2-DE gels, digested using trypsin and analysed by MALDI-TOF/TOF MS. The strategies used to identify these proteins involved a Mascot database search, PEAKS DB search and SPIDER homology search. The specific objectives of this thesis were to:

- Isolate at least 150 proteins extracted from gill tissue of the greenshell mussel *Perna canaliculus*
- Collect MS and MS/MS spectra from these proteins using MALDI-TOF/TOF MS
- Identify these proteins using a combined MS and MS/MS search against the Mollusca and NCBInr protein database using Mascot
- Make new identifications and confirm existing identifications using PEAKS DB and SPIDER
- Make new identifications by searching an Invertebrate EST database using Mascot
- Identify the challenges faced and how they could be overcome

2. Materials and Methods

2.1 Sample preparation

Gill tissue from *Perna canaliculus* was homogenised in 500 µl of ice-cold lysis buffer (30 mM Tris-Cl pH 8.8, 7 M urea, 2 M thiourea, 4% w/v CHAPS) to extract proteins. The homogenised gill extract was then transferred to a 1.5 ml centrifuge tube and left on ice for 30 min. The homogenate was then centrifuged at 10000 g for 5 min at room temperature and the supernatant transferred to a clean 1.5 ml centrifuge tube. Gill protein samples were then stored at -20 °C for later use.

2.2 Bradford protein assay

Protein assays were performed in a sterile 96-well polystyrene flat-bottom plate (Corning, New York, 16510035). In each well, 1 µl of protein sample was added to 200 µl of a 1:5 dilution of Bio-Rad protein assay solution (Bio-Rad, Hercules, California). Coomassie G-250, a component of Bio-Rad binds to aromatic and basic amino acid residues to induce a colour change. Absorbance was measured at 595 nm using an EnSpire 2300 Multilabel Plate Reader (PerkinElmer, Waltham, Massachusetts), which was compared to a standard curve containing 0, 2, 4, 6, 8 and 10 µg samples of BSA.

2.3 Two-dimensional gel electrophoresis

2.3.1 First dimension

Protein samples were separated in the first dimension using 7 cm Immobiline™ DryStrip Gels (GE Healthcare, Uppsala, Sweden) with a pH gradient of 4-7 and 6-11. IPG strip 4-7 was rehydrated in a reswelling tray using 300 µg of protein sample made up to 125 µl with IPG buffer (GE Healthcare). The IPG strip 6-11 was rehydrated using only 125 µl of IPG buffer, but was cuploaded immediately prior to IEF using 300 µg of protein sample made up to 125 µl with IPG buffer. PlusOne DryStrip Cover Fluid (GE Healthcare) was applied to each IPG strip and left overnight.

IEF was carried out using the Multiphor II electrophoresis system (Pharmacia Biotech). The Multiphor II system was configured with the temperature set at 20 °C, power at 5 W and a current of 2 mA. The IPG strip 4-7 was focused using three cycles: 200 V for 1 min; 3500 V for 1.5 h and another cycle of 3500 V for 1.5 h. The IPG strip 6-11 was run using slightly different conditions: of 200 V for 1 min; 3500 V for 1.05 h and 3500 V for 1.5 h.

Reduction and alkylation of IPG strips was carried out using 1% dithiothreitol, then 2.5% iodoacetamide, dissolved in 2 ml of equilibration buffer (50 mM Tris-Cl pH 8.8, 6 M urea, 30% glycerol, 2% w/v SDS) for 15 min. Dithiothreitol reduces disulphide bonds and iodoacetamide alkylates cysteine residues to prevent them from re-oxidising.

2.3.2 Second dimension

NuPAGE 4-12% Bis-Tris 1.0 mm gels (Invitrogen, Ontario, Canada) were loaded into a Novex mini-cell gel electrophoresis tank and covered with 20 x NuPAGE SDS running buffer diluted in a 1:20 ratio (Invitrogen, NP0001). The IPG strips were then placed into each gel, along with a Benchmark pre-stained protein ladder standard (Invitrogen, 10748-010). To keep proteins in a reduced state, 500 µl of antioxidant (Invitrogen) was added to the upper cathode (-) electrode buffer. The gels were run under the conditions of 400 V and 200 mA for 50 min.

Gels were fixed overnight in 50% ethanol and 3% phosphoric acid, washed in triple distilled water, and then placed in pre-staining solution [34% methanol, 17% ammonium sulphate and 3% phosphoric acid] for 1 h. Coomassie G-250 was then added to the staining solution and the gels were left to stain for 3 days. Gels were then washed in triple distilled water and scanned using a Molecular Dynamics scanner. ImageQuant 5.2 software was used to visualise the gel.

2.4 In-gel trypsin digestion

Protein spots were removed from the gel using a OneTouch Plus Spot Picker and 1.5 mm tips (Gel Company, San Francisco, California). Each gel piece was placed into 50 µl of triple distilled water in a 96-well polypropylene v-bottom plate (BD Biosciences). Digestion was

carried out using an ETTAN Digester (Amersham Biosciences). Four cycles of destaining were carried out by immersing gel pieces in 100 µl of a 50 % methanol solution containing 50 mM ammonium bicarbonate for 30 min. Gel pieces were then left to dry at room temperature for 1 h before adding trypsin.

Trypsin aliquots were made by suspending 25 µg of lyophilized modified sequencing grade trypsin (Roche, Mannheim, Germany) in 500 µl of triple distilled water. Ten 50 µl aliquots were made, each containing 2.5 µg of trypsin. Each aliquot was dried and resuspended in 500 µl of freshly prepared 20 mM ammonium bicarbonate. For each gel piece, 10 µl or 50 ng was added and left to digest the proteins at room temperature for 5 h. Peptides were extracted using 3 cycles of 35 µl of a 50 % ACN, 0.1 % TFA solution, which were transferred to a new v-shaped 96-well polypropylene plate. The peptide-containing solution was left to dry for 2 days.

2.5 MALDI-TOF/TOF mass spectrometry

A matrix was prepared by adding 10 mg of CHCA to 1 ml of ACN and 0.1 % TFA (1:1 v/v), briefly vortexed and centrifuged at 10,000 g for 5 min. The supernatant was then transferred to a clean tube for use. Tryptic peptides were resuspended in 1 µl of a CHCA matrix and spotted onto a 384 Opti-TOF 123 x 81 mm MALDI plate and left to crystallise.

The MALDI plate was loaded into an AB SCIEX MALDI-TOF/TOF 5800 mass spectrometer and left for 30 min for pressures to equilibrate. The m/z ratio of precursor ions was acquired in MS mode using a reflector positive ion method and a 355 nm diode pulse laser. TOF/TOF™ Series Explorer™ 4.0 software was used to set up the method using the following settings: mass range of 800-4000 Da and a focus mass of 1500 Da; continuous stage motion with a velocity of 600 µm and 200 shots per spectrum, with the first 10 shots discarded and a laser intensity of 3510 and pulse rate of 400 Hz. A processing method was also used to specify the criteria surrounding the collection of spectra, which required a minimum S/N ratio of 15, local noise window of 50 and a cluster area S/N optimisation of 5.

The MALDI TOF/TOF mass spectrometer was externally calibrated using a TOF/TOF calibration mixture made specifically for TOF/TOF instruments (AB SCIEX, Framingham, Massachusetts). This calibration mixture contains peptides with a known m/z ratio and

includes: des-arg1-bradykinin 904.4680 m/z; angiotensin I 1296.6850 m/z; glu1-fibrinopeptidase 1570.6770 m/z; ACTH peptides 2093.0870 m/z, 2465.1990 m/z and 3657.9294 m/z. The criteria set used for calibration was a minimum S/N ratio of 15, with a mass tolerance of +/- 0.1 m/z and a minimum of 3 peaks required to match. An interpretation method was used to specify the criteria for peptides entering MS/MS. Precursor ions required an S/N ratio greater than 20 and needed to be within the mass range of 800-4000 Da. Fifteen of the strongest precursor ions that met these requirements were selected for MS/MS.

An exclusion list (Table 1) was also used to prevent common contaminants and interference spectra from entering MS/MS. Trypsin and matrix fragments are common contaminants, while polyethylene glycol is a product of the materials used. Other contaminating peptides excluded were peptides found in the calibration mix. Interference spectra Masses were observed from a negative control with no protein. Their origin is unknown. Adducts with masses of 21.982 and 37.956 Da were also excluded.

MS/MS was carried out using the positive ion 1KV operating mode with CID on. Low power CID was used along with metastable suppressor. Each sub-spectrum which passed acceptance was accumulated, with acceptance criteria requiring an S/N ratio greater than 4. Stop conditions were initiated when 5 sub-spectra passed acceptance. The stage mode used was a continuous stage motion at a velocity of 1200 μm. Laser intensity was set at 4650, with 100 shots per spectrum allowed and a pulse rate of 1000 Hz. The processing method used specified a minimum S/N ratio of 10, local noise window of 250 and a cluster area S/N optimisation of 10. MS/MS mode was externally calibrated using the angiotensin I precursor ion with a m/z of 1296.6850. Acceptance required a minimum S/N ratio of 1, mass tolerance within +/- 0.1 m/z, with a minimum of 4 peaks to match.

Table 1.

List of m/z ratios excluded from MS/MS analysis. Contains known contaminants and interference spectra from an unknown origin.

Contaminating peptides			Interference Spectra	
m/z	Name	Tolerance (+/-)	m/z	Tolerance (+/-)
659.384	Trypsin	0.03	1007.646	0.03
805.417	Trypsin	0.03	1017.66	0.03
861.06	CHCA	0.1	1019.659	0.03
877	Polyethylene glycol	0.1	1033.683	0.03
906.505	Trypsin	0.03	1051.684	0.03
1020.503	Trypsin	0.03	1131.684	0.03
1153.574	Trypsin	0.03	1133.688	0.03
1175.523	Trypsin	0.03	1151.681	0.03
1296.68	Angiotensin 1	0.03	1165.704	0.03
1433.721	Trypsin	0.03	1265.716	0.03
1493.599	Trypsin	0.03	1279.725	0.03
1676.777	Trypsin	0.03	1300.83	0.03
1774.851	Trypsin	0.03	1302.83	0.03
2093.08	ACTH (clip 1-17)	0.03	1334.837	0.03
2163.057	Trypsin	0.03	1416.853	0.03
2193.003	Trypsin	0.03	1434.855	0.03
2193.995	Trypsin	0.03	1448.866	0.03
2273.16	Trypsin	0.03	1548.884	0.03
2289.155	Trypsin	0.03	1562.901	0.03
2305.15	Trypsin	0.03	1618.007	0.03
2465.19	ACTH (clip 18-39)	0.03	1718.019	0.03
2514.339	Trypsin	0.03	1732.034	0.03
2550.233	Trypsin	0.03	1901.177	0.03
2612.181	Trypsin	0.03	2015.207	0.03
2613.35	Trypsin	0.03	2162.99	0.03
3211.475	Trypsin	0.03	2289.084	0.03

2.6 Mascot database search

Peak lists were transferred to ProteinPilot™ 3.0 as DAT files using the Peaks to Mascot functionality of TOF/TOF Series Explorer. The Spot-Based MS/MS functionality of ProteinPilot™ was used for carrying out the Mascot search, which used the following search parameters: trypsin as the enzyme, with a maximum of one missed cleavage; carbamidomethylation of cysteine was as a fixed modification and oxidation of methionine as a variable modification; MALDI-TOF/TOF as the instrument of choice, along with monoisotopic mass values with a +1 charge and mass tolerance values were set at ± 50 ppm for peptide masses and ± 0.05 Da for fragment ions. A search was carried out against both a Mollusca and NCBInr protein database. The Mollusca database is made up of 58,900 protein sequences, while the NCBInr database contains 9,054,090 sequences. Both were downloaded in FASTA format from NCBI (http://www.ncbi.nlm.nih.gov).

2.7 PEAKS Studio 5.3

ABI 4700 Data Extractor (Bioinformatics Solutions, Waterloo, Canada) was used to obtain MS/MS peak lists in the form of PKL files from TOF/TOF Series Explorer. PEAKS studio 5.3 (Bioinformatics Solutions) refined these peak lists using the following parameters: correct precursor mass; +1 charge state; recommended quality filter of 0.65; peak centroiding, charge deconvolution and deisotoping. PEAKS *de novo* sequencing was performed selecting carbamidomethylation of cysteine as a fixed modification and oxidation of methionine as a variable modification. The mass tolerance was set at ± 50 ppm for peptide masses and ± 0.8 Da for fragment ions, while trypsin was also specified. An average local confidence value of 30 or greater was applied to filter *de novo* sequences.

De novo sequences with an average local confidence value of 50 or greater were searched against both a Mollusca and NCBInr database using PEAKS DB. The search was carried out for each individual protein spot by specifying trypsin as the enzyme used with a maximum of one missed cleavage. Carbamidomethylation of cysteine was selected as a fixed modification and oxidation of methionine as a variable modification. Monoisotopic mass values were also selected along with a peptide mass tolerance of ± 50 ppm and fragment ion tolerance of ± 0.8 Da. A decoy search was also performed. The Mollusca database is

made up of 58,900 protein sequences, while the NCBInr database contains 9,054,090 sequences. Both were downloaded in FASTA format from NCBI (http://www.ncbi.nlm.nih.gov)

A SPIDER homology search was carried out using *de novo* sequences with an average local confidence of 50 or greater. Search criteria specified carbamidomethylation of cysteine as a fixed modification and oxidation of methionine as a variable modification. Fragment ion tolerance was also set to 0.8Da. The search was only carried out against the Mollusca database. Leucine was selected as being equal to isoleucine and lysine as the equivalent to glutamine.

2.8 Mascot MS/MS ions search

Unidentified proteins were searched against an Invertebrate EST database using the online MS/MS ion search functionality of Mascot (http://www.matrixscience.com).

The search was performed by uploading raw PKL files containing MS/MS peaks lists and specified the following criteria: trypsin as the enzyme with a maximum of one missed cleavage; carbamidomethylation of cysteine as a fixed modification and oxidation of methionine as a variable modification; MALDI-TOF/TOF was selected as the instrument of choice, along with monoisotopic mass values with a +1 charge. Mass tolerance was set at ± 50 ppm for peptide masses and ± 0.8 Da for fragment ions. A decoy search was also performed.

3. Results

3.1 Two-dimensional gel electrophoresis

A 2-DE approach was used to separate proteins from the gill proteome of the greenshell mussel *Perna canaliculus*. In total, approximately 650 protein spots were resolved: 500 using the 4-7 gel (Figure 1a) and 150 using the 6-11 gel (Figure 1b). Protein amounts varied from high to low abundance and some were found to be adjoining or adjacent to other protein spots. One-hundred and fifty five protein spots of high to mid abundance were excised from the gel for further analysis. These were selected in a manner to provide a comprehensive coverage of proteins with a different pl and Mr. Both gels were scanned using a Molecular Dynamics scanner and visualised using ImageQuant 5.2 software.

For the 4-7 gel, proteins resolved best within the 6-7 pl range, while in the 5-6 pl range overabundant proteins were prominent. These overabundant proteins display vertical streaking and can be seen to impede low abundance proteins located underneath. Overabundant proteins were also observed in the poorly resolved 4-5 pl region for proteins with a Mr less than 40 kDa. For the 6-11 gel, its resolving ability gradually diminished at higher pl values and low abundance proteins were a recurring theme. Spot trains were also observed.

2-DE separation of gill proteins from *Perna canaliculus* using a 4-7 pl gradient

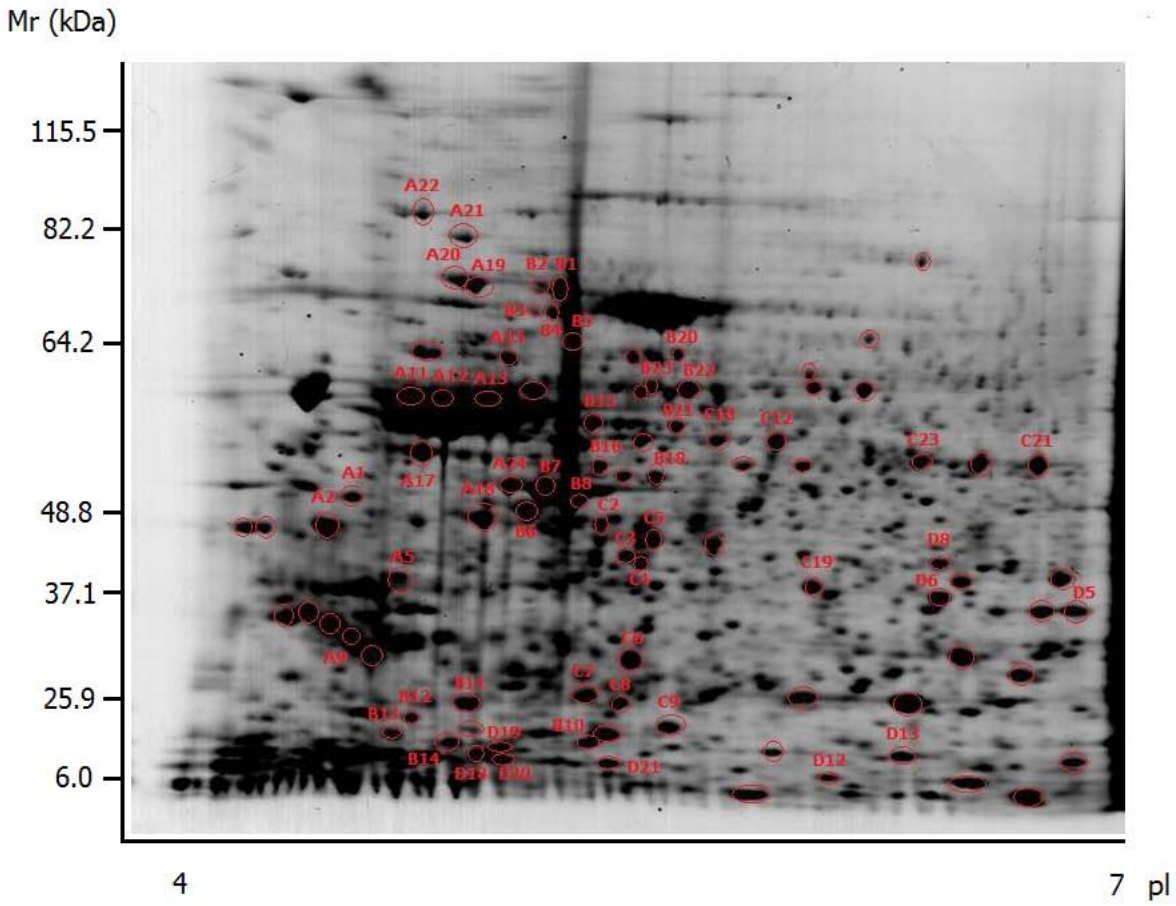


Figure 1 (a)

300 µg gill protein samples from the greenshell mussel *Perna canaliculus* were separated using a 4-7 pl gradient. Proteins were extracted from gill tissue, separated using 2-DE gel electrophoresis and stained using Coomaisse G-250. Gels were then scanned and visualised using Molecular Dynamics scanner and ImageQuant 5.2 software. (red circles indicate the protein spots excised, while numbered spots are those that were identified)

2-DE separation of gill proteins from *Perna canaliculus* using a 6-11 pl gradient

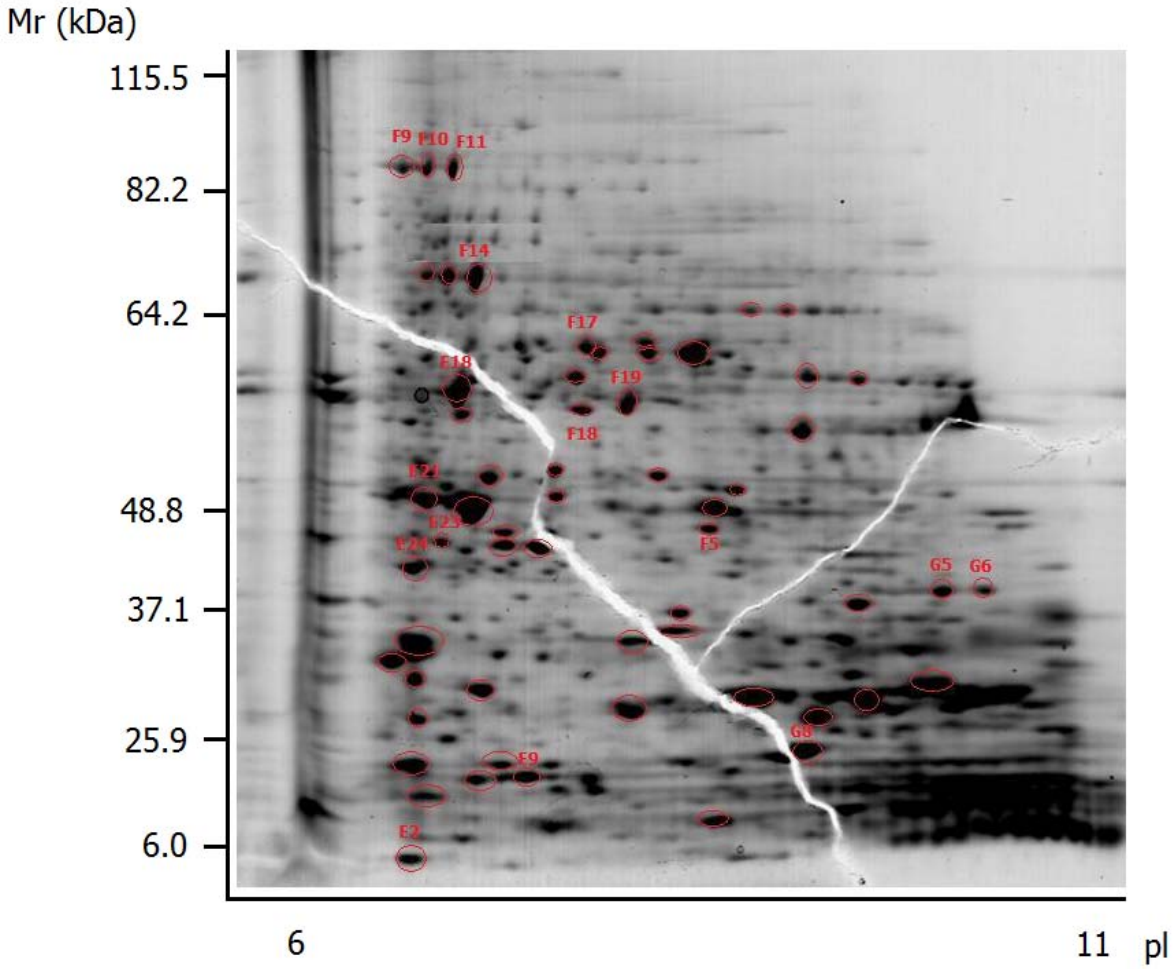


Figure 1 (b)

300 µg gill protein samples from the greenshell mussel *Perna canaliculus* were separated using a 6-11 pl gradient. Proteins were extracted from gill tissue, separated using 2-DE gel electrophoresis and stained using Coomaisse G-250. Gels were then scanned and visualised using Molecular Dynamics scanner and ImageQuant 5.2 software. (red circles indicate the protein spots excised, while numbered spots are those that were identified)

3.2 Mascot database search

3.2.1 Peptide mass fingerprinting

A combined MS and MS/MS ion search was carried out using Mascot against both the Mollusca and NCBInr database. From a search involving 155 proteins, 57 were identified by PMF (Table 2). Forty-one of these proteins were identified by searching the Mollusca database, while the other 16 were identified from searching the NCBInr database. Five identifications for protein spots A17, A18, C7, E24 and F17 were subsequently removed since they matched to hypothetical proteins only. A17, A18, C7 and E24 were matched using the NCBInr database and E24 the Mollusca database.

Most of the proteins identified were the cytoskeletal proteins actin and tubulin. Sixteen proteins were identified as actin from the protein spots A24-B8, B11, C4-C8, D18 and D20, which included beta-actin and cytoplasmic actin. Alpha and beta tubulin chains also accounted for 11 identifications from the protein spots A9-A13, B10, B13-C3, C23 and D19. Actin and tubulin also produced some of the highest Mascot scores of 676 and 619, respectively. This was well in excess of the significance threshold of 60. Cytoskeleton-related proteins were also identified. Tropomyosin was identified from protein spot A2 and produced a Mascot score of 212, while axonemal dynein was identified from protein spot G5 and produced a score of 454.

Other noticeable protein groups identified were stress-response proteins and those involved with protein biosynthesis. The stress-response proteins include heat shock protein 60 (spot A15) and 90 (spot A21 and A22), as well as the 78 kDa glucose regulated protein (spot A19 and A20). The antioxidant enzymes superoxide dismutase (D13) and peroxiredoxin 4 (D5) were also identified. All of the proteins involved in protein biosynthesis were identified from the NCBInr database search, except for the 40S ribosomal protein (A1). These proteins include the eukaryotic translation initiation factor 5A (C9) and elongation factor 2 (F9-11). All other identifications had an assortment of functions.

Figure 2a-f presents the MS spectra for six protein spots A1, A12, B6, C12, E9 and G5. These show peptides matching the most abundant peaks, except in the case for spot E9. In this

case matches were made to the 3rd and 4th most abundant peak. MS spectra for all proteins identified by PMF are included in Appendix 1.

3.2.2 MS/MS ions search against Mollusca or NCBInr database

An MS/MS ion search confirmed the identifications made by PMF, but identified an additional four proteins (Table 3). These include heat shock protein 70 (spot B1 and B2), mitochondrial mortalin 2 (also spot B2) and GTP-binding protein beta subunit (spot C5).

The MS/MS ion search also made seven single-peptide identifications (Table 4). Heat shock protein 90 (A22), threonine dehydrogenase (F5) and serine hydroxymethyltransferase (F18) were all identified from searching the NCBInr database, while beta-tubulin (A9), actin (B11), malate dehydrogenase (D8) and peptidyl prolyl cis-trans isomerase A (G8) were identified from searching the Mollusca database. Figure 3-6 shows MS/MS spectra with accompanying fragment ion assignments for protein spots A9, B11, D8 and G8. Y-series ions are a dominating feature among the assigned fragment ions, except for protein spot G8 where there is a higher representation of b-series ions.

3.2.3 MS/MS ions search against Invertebrate EST database

A MS/MS ions search was carried out for unidentified proteins against the Invertebrate EST database using Mascot. This search strategy identified five new proteins. Protein spots A11, C6 and E18 were used as positive controls. New identifications were given as tropomyosin (A5, Gelsolin (B16), T-complex protein 1 beta (B20), Tektin (B21), Enkurin (G6). B16 produced the highest score of 120, while B20 only just sneeked past the significance threshold with a score of 59. The significance threshold was 57. All were identified by a single peptide. Figure 13 to 16 shows MS/MS spectra and fragment ion assignments, all predominating with y ion assisnments.

Table 2. Summary results for proteins producing a significant score from a Mascot search.

Spot ID[a]	Accession no.	Protein score*	Expect	Theoretical/Observed *M*r (kDa)	No. of matching peptide masses/searched	Sequence coverage (%)[b]	Protein description
A1	gi|229891605	171	4.70e-13	33.7/50.1	6/13	30.9	40S ribosomal protein SA
A2	gi|9954251	212	3.70e-17	32.8/46.7	5/16	30.6	Tropomyosin
A9	gi|1066143	80	6.3e-4	38.7/32.3	1/12	5.3	Beta-tubulin
A11	gi|194068375	518	9.30e-48	50.4/60.2	14/21	39.5	Beta-tubulin
A12	gi|194068375	589	7.40e-55	50.4/60.2	15/21	41.9	Beta-tubulin
A13	gi|1174593	598	9.3e-56	50.9/60.2	13/16	43.1	Tubulin alpha-2/alpha-4 chain
A15	gi|223954136[n]	215	2.9e-15	61.1/63.8	2/3	7.3	Heat shock protein 60
A19	gi|46359618	163	3e-12	73.1/72.6	10/17	19.8	78 kDa glucose regulated protein
A20	gi|46359618	132	3.70e-9	73.1/73.2	4/11	8.6	78 kDa glucose regulated protein
A21	gi|153793258	95	2.00e-5	83.4/80.6	3/13	4.4	Heat shock protein 90
A22	gi|108760025[n]	102	5.70e-4	73.3/87.4	4/13	7.3	Heat shock protein 90
A24	gi|224305	306	1.50e-26	41.8/51.2	4/6	9.6	Actin
B1	gi|89255272	411	4.70e-37	41.3/74.2	7/17	25.6	Cytoplasmic actin
B2	gi|224305	300	5.90e-26	41.8/74.2	5/15	11.3	Actin
B3	gi|224305	220	5.9e-18	41.8/70.8	5/17	11.3	Actin
B4	gi|89255272	375	1.90e-33	41.3/70.8	6/8	23.5	Cytoplasmic actin
B5	gi|159507454	303	3.00e-26	42.1/65.7	7/12	26.6	Beta-actin
B6	gi|89255272	676	1.50e-63	41.3/49.7	9/13	31.8	Cytoplasmic actin
B7	gi|159507454	431	4.7e-39	42.1/41.6	7/13	26.6	Beta-actin
B8	gi|159507454	495	1.90e-45	42.1/51.7	8/19	28.7	Beta-actin
B10	gi|1174593	139	7.40e-10	50.9/14.1	2/19	8.0	Tubulin alpha-2/alpha-4 chain
B11	gi|42560365	91	4.4e-5	22.9/25.4	2/10	14.5	Actin
B12	gi|89268290[n]	110	9.10e-5	19.8/21.1	3/5	22.1	Myosin regulatory light chain 2
B13	gi|53801335	619	7.40e-58	42.3/16.1	9/17	35.4	Beta-tubulin

B15	gi|1174593	299	7.40e-26	50.9/56.4	7/8	25.0	Tubulin alpha-2/alpha-4 chain
C2	gi|1174593	497	1.2e-45	50.9/47.2	7/11	25.0	Tubulin alpha-2/alpha-4 chain
C3	gi|1335661	612	3.70e-57	50.1/42.9	11/19	33.6	Beta tubulin
C4	gi|483321	154	2.30e-11	42.2/44.1	3/11	12.0	Actin
C5	gi|159507454	225	1.50e-18	42.1/45.9	5/12	21.0	Beta-actin
C6	gi|159507454	224	2.30e-18	42.1/30.4	4/9	16.2	Beta-actin
C8	gi|159507454	194	2.30e-15	42.1/25.8	5/12	19.9	Beta-actin
C9	gi|113171152[n]	161	7.2e-10	17.4/23.2	2/8	15.9	Eukaryotic translation initiation factor 5A
C10	gi|14423688[n]	158	1.40e-9	48.0/55.6	3/10	10.1	Enolase 1
C12	gi|1169529[n]	186	2.30e-12	43.2/55.6	4/14	13.2	Enolase 1
C23	gi|1174593	115	1.90e-7	50.9/53.9	3/12	11.1	Tubulin alpha-2/alpha-4 chain
D5	gi|209171293	78	9.10e-4	19.3/35.1	2/6	14.6	Peroxiredoxin 4 variant precursor
D8	gi|6746611	149	7.40e-11	36.6/42.3	3/12	15.8	Malate dehydrogenase precursor
D13	gi|215263232	107	1.20e-6	15.9/16.1	2/10	19.1	Superoxide dismutase
D18	gi|2642634	110	5.90e-7	18.1/14.4	3/6	23.6	Actin
D19	gi|1174593	195	1.90e-15	50.9/15.8	2/7	8.0	Tubulin alpha-2/alpha-4 chain
D20	gi|2642634	186	1.50e-14	18.1/11.2	3/9	23.6	Actin
E2	gi|46359622	347	1.90e-33	77.0/6.0	5/7	8.9	Polyubiquitin
E9	gi|126697388	89	7.90e-5	18.9/18.4	2/12	15.5	Nucleoside diphosphate kinase B
E18	gi|116008297	152	3.70e-11	59.8/57.6	5/13	11.4	Mitochondrial H+ ATPase alpha subunit
F5	gi|112982820[n]	106	2.30e-4	39.3/45.8	1/10	4.7	L-threonine dehydrogenase
F9	gi|16554298[n]	99	1.20e-3	94.4/91.2	2/8	3.0	Elongation factor 2
F10	gi|16554298[n]	85	2.70e-2	94.4/91.4	2/7	3.0	Elongation factor 2
F11	gi|16554298[n]	106	2.30e-4	94.4/91.3	3/13	3.9	Elongation factor 2
F18	gi|66816019[n]	98	1.50e-3	50.8/54.1	2/16	4.8	Serine hydroxymethyltransferase
F19	gi|28564385[n]	161	7.2e-10	27.7/55.4	2/9	9.6	GND1

G5	gi|126697474	454	2.30e-41	29.2/41.9	10/17	46.9	Axonemal dynein light chain p33
G8	gi|289064181	78	9.3e-4	17.7/24.7	1/17	9.1	Peptidyl prolyl cis-trans isomerase A

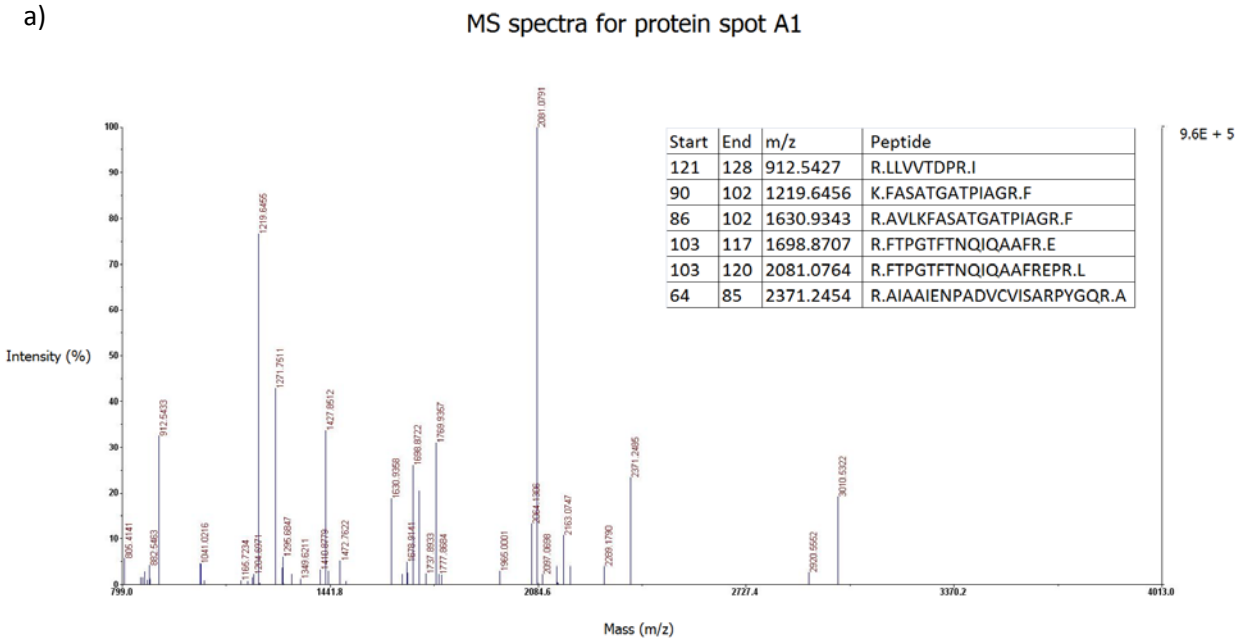
[a] refers to protein spots from figure 1a and 1b
[b] calculated by diving the number of amino acids of peptides identified by MS by the protein amino acid length
[n] refers to identifications from searching the NCBInr database

* Protein scores greater than 60 were significant ($p<0.05$) for the Mollusca database search, scores greater than 82 were significant ($p<0.05$) for the NCBInr database search.

The search was carried out using the MS/MS spot-based functionality as part of the ProteinPilot 3.0 software package. The protein score is derived from the ions scores as a non-probabilistic basis for ranking protein hits and is represented by the equation -10*Log(P), where (P) is the probability that the observed match is a random event. The Mascot search was carried out specifying trypsin as the enzyme, with a maximum of one missed cleavage allowed. Carbamidomethylation of cysteine was selected as a fixed modification and oxidation of methionine as a variable modification. MALDI-TOF/TOF was selected as the instrument used along with a +1 charge for monoisotopic mass values. Mass tolerance was set at ± 50 ppm for peptide masses and ± 0.05 Da for fragment ions. Database searches were carried out against both the Mollusca and NCBInr protein sequence databases. The Mollusca database contained 58,900 sequences and NCBInr had 9,054,090 sequences.

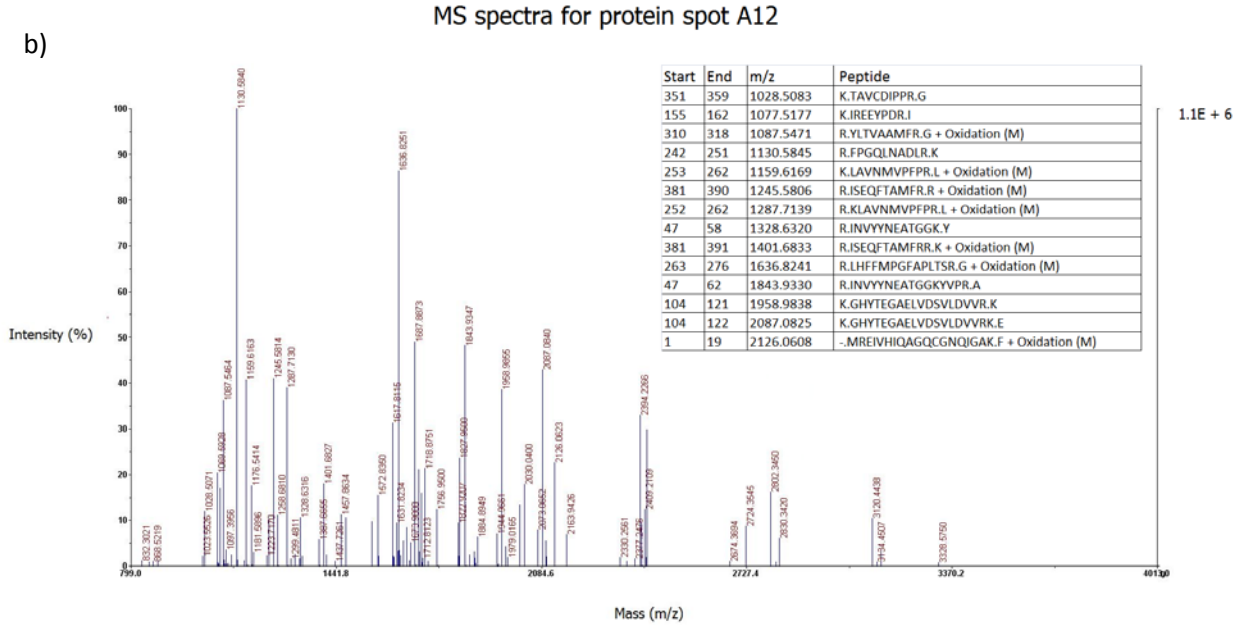
Figure 2. MS spectrum for selected proteins (a-f) accompanied with a table showing peptide mass assignments.

a)



Start	End	m/z	Peptide
121	128	912.5427	R.LLVVTDPR.I
90	102	1219.6456	K.FASATGATPIAGR.F
86	102	1630.9343	R.AVLKFASATGATPIAGR.F
103	117	1698.8707	R.FTPGTFTNQIQAAFR.E
103	120	2081.0764	R.FTPGTFTNQIQAAFREPR.L
64	85	2371.2454	R.AIAAIENPADVCVISARPYGQR.A

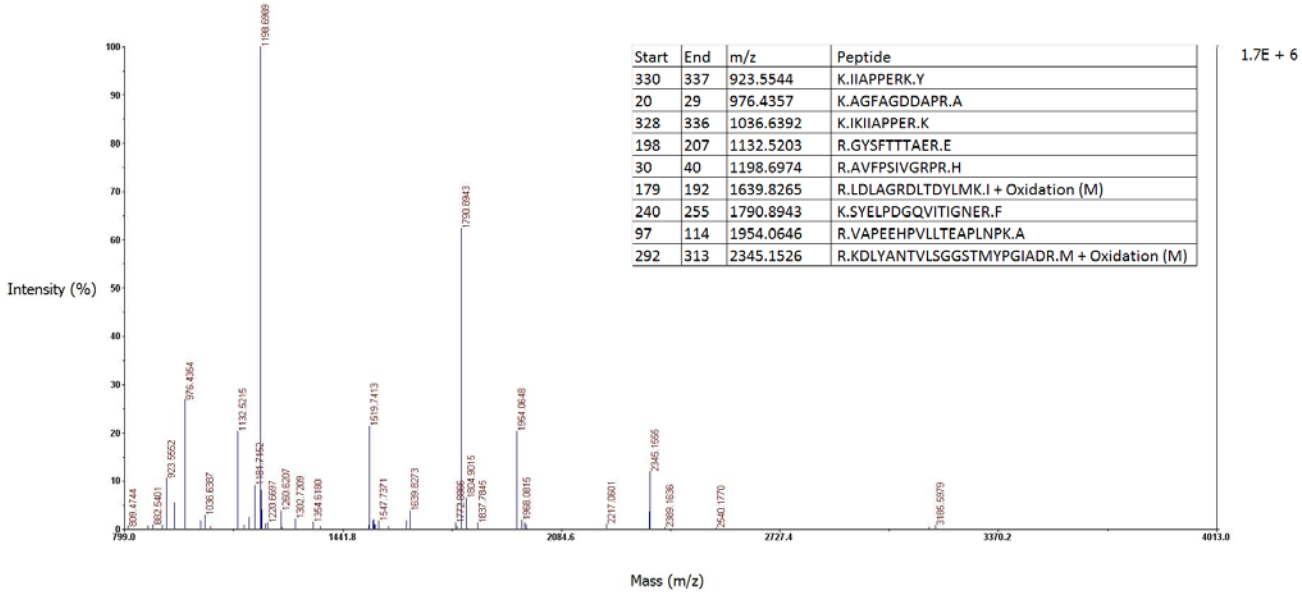
b)



Start	End	m/z	Peptide
351	359	1028.5083	K.TAVCDIPPR.G
155	162	1077.5177	K.IREEYPDR.I
310	318	1087.5471	R.YLTVAAMFR.G + Oxidation (M)
242	251	1130.5845	R.FPGQLNADLR.K
253	262	1159.6169	K.LAVNMVPFPR.L + Oxidation (M)
381	390	1245.5806	R.ISEQFTAMFR.R + Oxidation (M)
252	262	1287.7139	R.KLAVNMVPFPR.L + Oxidation (M)
47	58	1328.6320	R.INVYYNEATGGK.Y
381	391	1401.6833	R.ISEQFTAMFRR.K + Oxidation (M)
263	276	1636.8241	R.LHFFMPGFAPLTSR.G + Oxidation (M)
47	62	1843.9330	R.INVYYNEATGGKYVPR.A
104	121	1958.9838	K.GHYTEGAELVDSVLDVVR.K
104	122	2087.0825	K.GHYTEGAELVDSVLDVVRK.E
1	19	2126.0608	-.MREIVHIQAGQCGNQIGAK.F + Oxidation (M)

c)

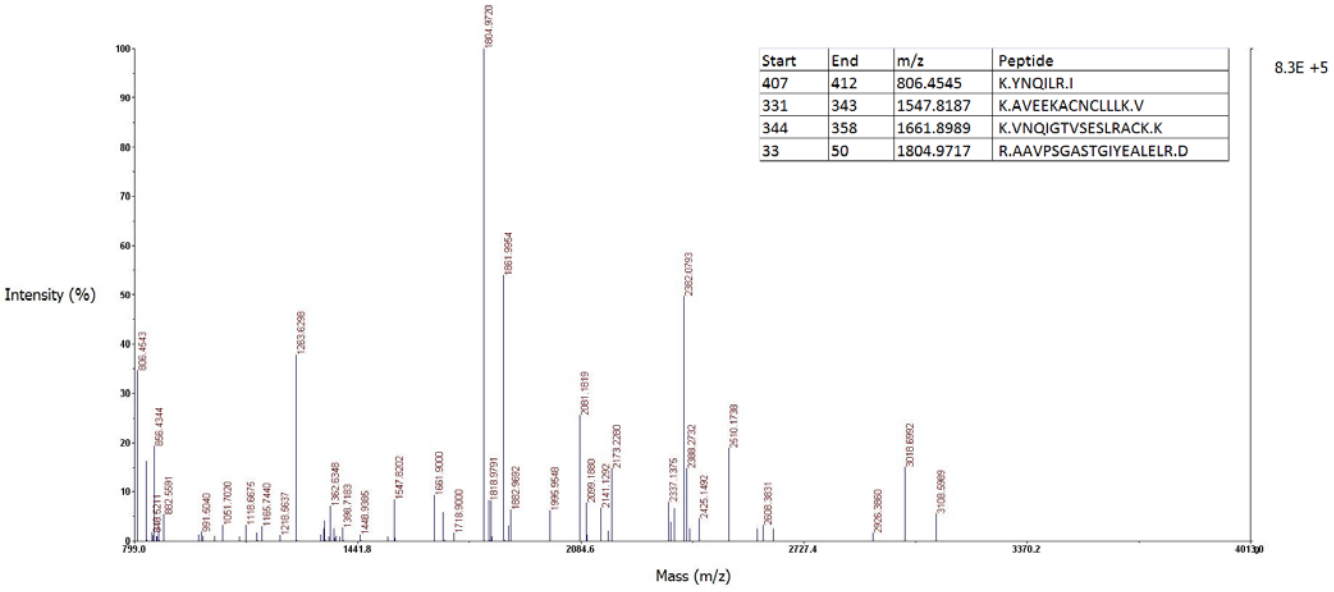
MS spectra for protein spot B6



Start	End	m/z	Peptide
330	337	923.5544	K.IIAPPERK.Y
20	29	976.4357	K.AGFAGDDAPR.A
328	336	1036.6392	K.IKIIAPPER.K
198	207	1132.5203	R.GYSFTTTAER.E
30	40	1198.6974	R.AVFPSIVGRPR.H
179	192	1639.8265	R.LDLAGRDLTDYLMK.I + Oxidation (M)
240	255	1790.8943	K.SYELPDGQVITIGNER.F
97	114	1954.0646	R.VAPEEHPVLLTEAPLNPK.A
292	313	2345.1526	R.KDLYANTVLSGGSTMYPGIADR.M + Oxidation (M)

d)

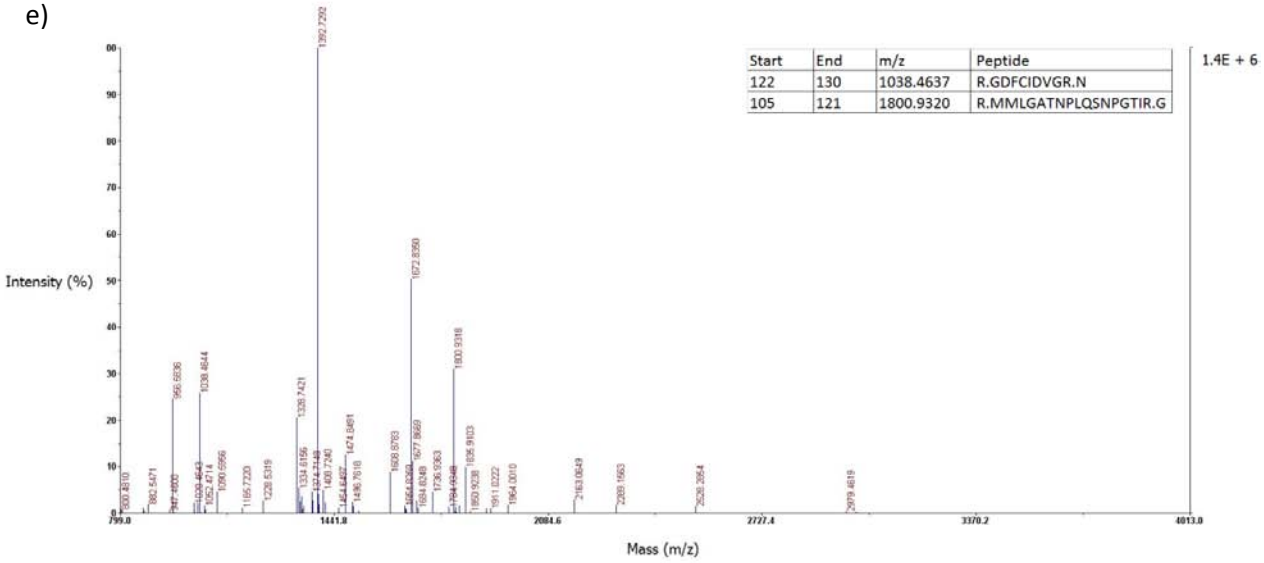
MS spectra for protein spot C12



Start	End	m/z	Peptide
407	412	806.4545	K.YNQILR.I
331	343	1547.8187	K.AVEEKACNCLLLK.V
344	358	1661.8989	K.VNQIGTVSESLRACK.K
33	50	1804.9717	R.AAVPSGASTGIYEALELR.D

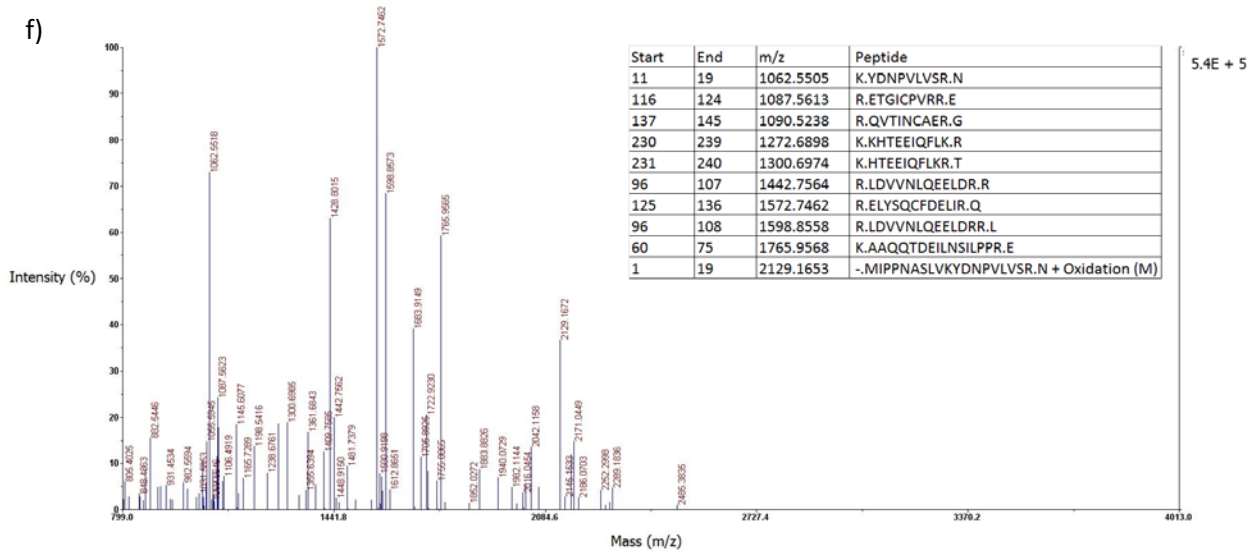
MS spectra for protein spot E9

e)



MS spectra for protein spot G5

f)



The raw MS spectrum (with annotated masses) can be matched to the table of peptide assignments for each protein identified. The table shows the list of peptides that were correctly matched for each protein and their theoretical masses. AB SCIEX Data Explorer was used to refine each spectrum allowing for a signal-to-noise threshold of 30, peak centroiding and deisotoping.

Table 3. Summary results for proteins identified by 2 or more peptides from a Mascot search.

Spot ID[a]	Accession no.[n]	Protein description	No. of unique peptides/ total peptides	Theoretical/Observed Mr (kDa)	Individual ions score*	m/z[c]	Sequence confirmed by MS/MS[d]
A1	gi|126362082	67 kDa laminin receptor precursor	5/5	33.7/50.1	28	1698.8707	R.FTPGTFTNQIQAAFR.E
					41	2081.0764	R.FTPGTFTNQIQAAFREPR.L
A2	gi|9954251	Tropomyosin	3/3	32.8/46.7	138	1738.8989	K.QIQEHEQEIQSLTR.K
					26	2511.1921	K.NIQTENDYDNCNTQLQDVQAK.Y
A11	gi|51860821	Beta-tubulin	6/9	38.7/32.3	24	1087.5609	R.YLTVAAMFR.G + oxidation (M)
					85	1130.6005	R.FPGQLNADLR.K
					89	1959.0177	K.GHYTEGAELVDSVLDVVR.K
					82	2087.1135	K.GHYTEGAELVDSVLDVVRK.E
A12	gi|194068375	Beta-tubulin	6/9	50.4/60.2	81	1130.5845	R.FPGQLNADLR.K
					54	1636.8241	R.LHFFMPGFAPLTSR.G + oxidation(M)
					50	1843.933	R.INVYYNEATGGKYVPR.A
					104	1958.9838	K.GHYTEGAELVDSVLDVVR.K
					100	2087.0825	K.GHYTEGAELVDSVLDVVRK.E
	gi|454315	Alpha-tubulin	2/2	50.4/60.2	90	1687.8859	R.AVFVDLEPTVVDEVR.T
					50	2415.2073	R.QLFHPEQLITGKEDAANNYAR.G
A13	gi|454315	Alpha-tubulin	2/7	50.9/60.2	53	1410.7616	R.QLFHPEQLITGK.E
					65	1457.8674	R.LIGQIVSSITASLR.F
					128	1687.8973	R.AVFVDLEPTVVDEVR.T
					30	1718.8839	R.NLDIERPTYTNLNR.L
					73	1756.9602	R.IHFPLATYAPVISAEK.A
					23	1824.983	K.VGINYQPPTVVPGGDLAK.V
					61	2415.2202	R.QLFHPEQLITGKEDAANNYAR.G

A15	gi|40647591[n]	Mitochondrial 60 kDa heat shock protein	0/2	61.1/63.8	96	1607.9154	R.AAVEEGIVPGGGVALIR.C
					92	2560.2605	K.LVQDVANNTNEEAGDGTTTATVLAR.T
A19	gi|46359618	78 kDa glucose regulated protein	2/3	73.1/72.6	57	1183.6555	K.FDLTGIPPAPR.G
A20	gi|46359618	78 kDa glucose regulated protein	2/3	73.1/73.2	24	1183.6732	K.FDLTGIPPAPR.G
					64	1788.0261	R.IINEPTAAAIAYGLDKK.E
A21	gi|38146757	Heat shock protein 90	2/2	83.4/80.6	27	815.5519	R.ALLFVPR.R
					53	1348.7318	K.HFSVEGQLEFR.A
A24	gi|224305	Actin	3/4	41.8/51.2	30	976.4874	K.AGFAGDDAPR.A
					69	1198.7535	R.AVFPSIVGRPR.H
					121	1790.9609	K.SYELPDGQVITIGNER.F
					33	1954.1224	R.VAPEEHPVLLTEAPLNPK.A
B1	gi|224305	Actin	2/5	41.3/74.2	56	976.4302	K.AGFAGDDAPR.A
					58	1132.5142	R.GYSFTTTAER.E
					66	1198.6913	R.AVFPSIVGRPR.H
					94	1790.8871	K.SYELPDGQVITIGNER.F
					71	1954.0602	R.VAPEEHPVLLTEAPLNPK.A
	gi|57635269	Heat shock protein 70	¾	41.3/74.2	62	1408.7806	K.AAVHEIVLVGGSTR.I
					80	1480.7443	R.ARFEELNADLFR.G
					110	1707.7119	K.STSGDTHLGGEDFDNR.M
B2	gi|224305	Actin	2/4	41.8/74.2	34	976.4408	K.AGFAGDDAPR.A
					48	1198.7042	R.AVFPSIVGRPR.H
					97	1790.9061	K.SYELPDGQVITIGNER.F
					77	1954.0745	R.VAPEEHPVLLTEAPLNPK.A
	gi|57635269	Heat shock protein 70	2/3	41.8/74.2	40	1408.7916	K.AAVHEIVLVGGSTR.I
					74	1480.7587	R.ARFEELNADLFR.G
					66	1707.7356	K.STSGDTHLGGEDFDNR.M
	gi|93009035	Mitochondrial mortalin 2	2/2	41.8/74.2	59	1242.68	K.DAGQISGLNVLR.V
					85	1680.844	K.NAVVTVPAYFNDSQR.Q

B4	gi|224305	Actin	2/4	41.3/70.8	54	976.4658	K.AGFAGDDAPR.A
					33	1198.7343	R.AVFPSIVGRPR.H
					121	1790.943	K.SYELPDGQVITIGNER.F
					82	1954.1161	R.VAPEEHPVLLTEAPLNPK.A
B5	gi|224305	Actin	2/4	42.1/65.7	51	1198.7051	R.AVFPSIVGRPR.H
					111	1790.9009	K.SYELPDGQVITIGNER.F
					43	1954.0709	R.VAPEEHPVLLTEAPLNPK.A
B6	gi|56693681	Actin ovestestis isoforms	2/9	41.3/49.7	30	923.5544	K.IIAPPERK.Y
					90	976.4357	K.AGFAGDDAPR.A
					48	1132.5203	R.GYSFTTTAER.E
					64	1198.6974	R.AVFPSIVGRPR.H
					127	1790.8943	K.SYELPDGQVITIGNER.F
					130	1954.0646	R.VAPEEHPVLLTEAPLNPK.A
					33	2345.1526	R.KDLYANTVLSGGSTMYPGIADR.M + oxidation (M)
B7	gi|224305	Actin	2/5	42.1/41.6	68	976.4434	K.AGFAGDDAPR.A
					68	1198.7085	R.AVFPSIVGRPR.H
					113	1790.9122	K.SYELPDGQVITIGNER.F
					89	1954.0863	R.VAPEEHPVLLTEAPLNPK.A

B8	gi|71148423	Actin	2/6	42.1/51.7	59	976.4349	K.AGFAGDDAPR.A
					64	1132.5198	R.GYSFTTTAER.E
					62	1198.6964	R.AVFPSIVGRPR.H
					118	1790.8879	K.SYELPDGQVITIGNER.F
					83	1954.0595	R.VAPEEHPVLLTEAPLNPK.A
B10	gi|454315	Alpha-tubulin	2/2	50.9/14.1	80	1687.9216	R.AVFVDLEPTVVDEVR.T
					46	2415.2529	R.QLFHPEQLITGKEDAANNYAR.G

B13	gi|1335661	Beta-tubulin	1/6	42.3/16.1	33	1077.5182	K.IREEYPDR.I
					77	1328.6338	R.INVYYNEATGGK.Y
					55	1617.8113	R.AVLVDLEPGTMDSVR.S + oxidation (M)
					111	1843.9386	R.INVYYNEATGGKYVPR.A
					151	1958.9896	K.GHYTEGAELVDSVLDVVR.K
					120	2087.0889	K.GHYTEGAELVDSVLDVVRK.E
B15	gi|454315	Alpha-tubulin	3/4	50.9/56.4	124	1687.9127	R.AVFVDLEPTVVDEVR.T
					28	1718.9032	R.NLDIERPTYTNLNR.L
					41	2415.2393	R.QLFHPEQLITGKEDAANNYAR.G
C2	gi|454315	Alpha-tubulin	2/6	50.9/47.2	31	1410.7849	R.QLFHPEQLITGK.E
					84	1457.8877	R.LIGQIVSSITASLR.F
					121	1687.9204	R.AVFVDLEPTVVDEVR.T
					38	1718.9098	R.NLDIERPTYTNLNR.L
					75	1756.9862	R.IHFPLATYAPVISAEK.A
					61	2415.2583	R.QLFHPEQLITGKEDAANNYAR.G
C3	gi|1335661	Beta-tubulin	6/10	50.1/42.9	32	1077.5211	K.IREEYPDR.I
					85	1130.5879	R.FPGQLNADLR.K
					25	1287.7179	R.KLAVNMVPFPR.L + oxidation (M)
					54	1636.8301	R.LHFFMPGFAPLTSR.G + oxidation(M)
					67	1843.9381	R.INVYYNEATGGKYVPR.A
					100	1958.9921	K.GHYTEGAELVDSVLDVVR.K
					65	2087.0889	K.GHYTEGAELVDSVLDVVRK.E
C4	gi|224305	Actin	1/2	42.2/44.1	97	1790.9117	K.SYELPDGQVITIGNER.F
					32	1954.0828	R.VAPEEHPVLLTEAPLNPK.A
C5	gi|224305	Actin	1/3	42.1/45.9	108	1790.8889	K.SYELPDGQVITIGNER.F
					67	1954.0559	R.VAPEEHPVLLTEAPLNPK.A
	gi|9508	GTP-binding protein beta subunit	3/3	42.1/45.9	86	1549.6874	R.ELPGHTGYLSCCR.F

C6	gi|159507454	Beta-actin	2/4	42.1/30.4	30	1132.5359	R.GYSFTTTAER.E
					29	1516.7094	K.QEYDESGPSIVHR.K
					114	1790.9054	K.SYELPDGQVITIGNER.F
C8	gi|483321	Actin	2/2	42.1/25.8	40	1516.7192	K.QEYDESGPSIVHR.K
					102	1790.9158	K.SYELPDGQVITIGNER.F
C9	gi|113171152[n]	Eukaryotic translation initiation factor 5A	2/2	17.4/23.2	112	1312.7814	K.VHLIGIDLFTGK.K
C10	gi|14423688	Enolase 1	2/2	48.0/55.6	120	1804.9769	R.AAVPSGASTGIYEALELR.D
					88	1862.0005	R.GNPTVEVDLTTDKGIFR.A
C12	gi|53830714[n]	Enolase 1	0/2	43.2/55.6	153	1804.9717	R.AAVPSGASTGIYEALELR.D
C23	gi|454315	Alpha-tubulin	2/2	50.9/53.9	74	1687.9701	R.AVFVDLEPTVVDEVR.T
D5	gi|13488586	Thioredoxin peroxidase BgTPx	2/2	19.3/35.1	44	1591.83	K.AYGVYLQDLGHSLR.G
D13	gi|215263232	Superoxide dismutase	2/2	15.9/16.1	37	1017.5862	R.LACGVIGISK.V
					48	2090.1191	R.TVVVHADIDDLGKGGHELSK.T
D18	gi|224305	Actin	2/2	18.1/14.4	49	1198.7405	R.AVFPSIVGRPR.H
D19	gi|454315	Alpha-tubulin	2/2	50.9/15.8	78	1687.9563	R.AVFVDLEPTVVDEVR.T
					98	2415.2942	R.QLFHPEQLITGKEDAANNYAR.G
D20	gi|224305	Actin	3/3	18.1/11.2	54	1198.7405	R.AVFPSIVGRPR.H
					81	1954.1276	R.VAPEEHPVLLTEAPLNPK.A
E2	gi|12240012	Ubiquitin	5/5	77.0/6.0	24	1039.5199	K.EGIPPDQQR.L
					32	1346.7521	R.LIFAGKQLEDGR.T
					92	1523.792	K.IQDKEGIPPDQQR.L
					89	2130.1709	R.TLSDYNIQKESTLHLVLR.L
E9	gi|126697388	Nucleoside diphosphate kinase B	2/2	18.9/18.4	45	1038.4637	R.GDFCIDVGR.N
					26	1800.932	R.MMLGATNPLQSNPGTIR.G

E18	gi|116008297	Mitochondrial H+ ATPase a subunit	2/1	59.8/57.6	31	1553.8033	R.EAYPGDVFYLHSR.L
					82	2408.2778	R.EVAAFAQFGSDLDQATQNLLNR.G
F9	gi|16554298[n]	Elongation factor 2	2/2	94.4/91.2	57	1785.8767	K.AYLPVNESFGFDSALR.A
F19	gi|28564155[n]	GND1	2/2	27.7/55.4	73	1123.681	R.LPANLLQAQR.D
					76	1577.8363	K.GILFVGSGVSGGEDGAR.Y
G5	gi|126697474	Axonemal dynein light chain p33	9/9	29.2/41.9	48	1062.5505	K.YDNPVLVSR.N
					28	1272.6898	K.KHTEEIQFLK.R
					63	1442.7564	R.LDVVNLQEELDR.R
					35	1572.7462	R.ELYSQCFDELIR.Q
					46	1598.8558	R.LDVVNLQEELDRR.L
					84	1765.9568	K.AAQQTDEILNSILPPR.E

[a] refers to protein spots from figure 1a and 1b
[c] all peptides have a +1 charge
[d] sequences were confirmed by collision-induced dissociation
[n] refers to identifications from searching the NCBInr database

The search was carried out using the MS/MS spot-based functionality as part of the ProteinPilot 3.0 software package. Only peptide sequences with a score above the identity threshold are displayed Individual ion scores are based on the equation -10*Log(P), where (P) is the probability that the observed match is a random event.

* Ions score greater than 23 were significant ($p<0.05$) for the Mollusca database search, while scores greater than 45 were significant ($p<0.05$) for the NCBInr database search.

Search conditions: the search was carried out specifying trypsin as the enzyme with a maximum of one missed cleavage allowed. Carbamidomethylation of cysteine was selected as a fixed modification and oxidation of methionine as a variable modification. MALDI-TOF/TOF was selected as the instrument used along with a +1 charge for monoisotopic mass values. Mass tolerance was set at ± 50 ppm for peptide masses and ± 0.05 Da for fragment ions. Database searches were carried out against both the Mollusca and NCBInr protein sequence databases. The Mollusca database contained 58,900 sequences and NCBInr had 9,054,090 sequences.

Table 4. Peptide summary results for proteins identified by a single peptide from a Mascot search.

Spot ID[a]	Accession no.[n]	Protein description	Ions score*	Expect	Error (ppm)[e]	m/z[c]	Sequence confirmed by MS/MS[d]
A9	gi|1066143	Beta-tublin	73	2.3e-007	-14.21	1958.9540	K.GHYTEGAELVDSVLDVVR.K
A22	gi|108760025[n]	Heat shock protein 90	80	8.4e-005	23.6	1499.8050	R.GVIDSDDLPLNVSR.E
B11	gi|71148423	Actin	71	2.3e-006	-3.92	1790.8849	K.SYELPDGQVITIGNER.F
D8	gi|6746611	Malate dehydrogenase precursor	121	7.4e-012	12.7	1334.6870	R.DDLFNTNAGIVR.D
F5	gi|241599280[n]	Threonine dehydrogenase	98	1.3e-006	2.74	1660.8739	R.LFVPSTIGAFGPDSPR.H
F18	gi|66816019[n]	Serine hydroxymethyltransferase	88	1.1e-005	25.9	1436.7751	K.GLELIASENFTSR.A
G8	gi|289064181	Peptidyl prolyl cis-trans isomerase A	70	8.7e-007	2.40	1630.8297	K.HVVFGNVVDGMDVVK.A + oxidation of methionine

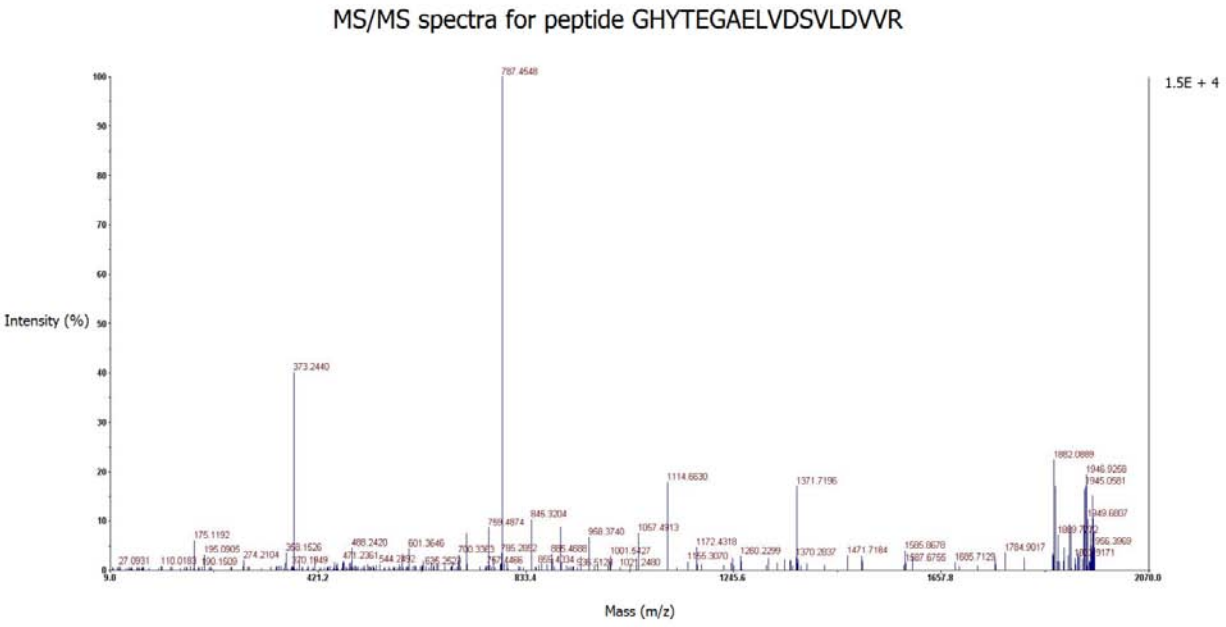
[a] refers to protein spots from figure 1a and 1b
[c] all peptides have a +1 charge
[d] sequences were confirmed by collision-induced dissociation
[e] parts per million (PPM)
[n] refers to identifications from searching the NCBInr database

The search was carried out through the MS/MS spot-based functionality as part of the ProteinPilot 3.0 software package. Individual ion scores are based on the equation -10*Log(P), where (P) is the probability that the observed match is a random event.

* Ions score greater than 23 were significant ($p<0.05$) for the Mollusca database search, scores greater than 45 were significant ($p<0.05$) for the NCBInr database search.

Figure 3

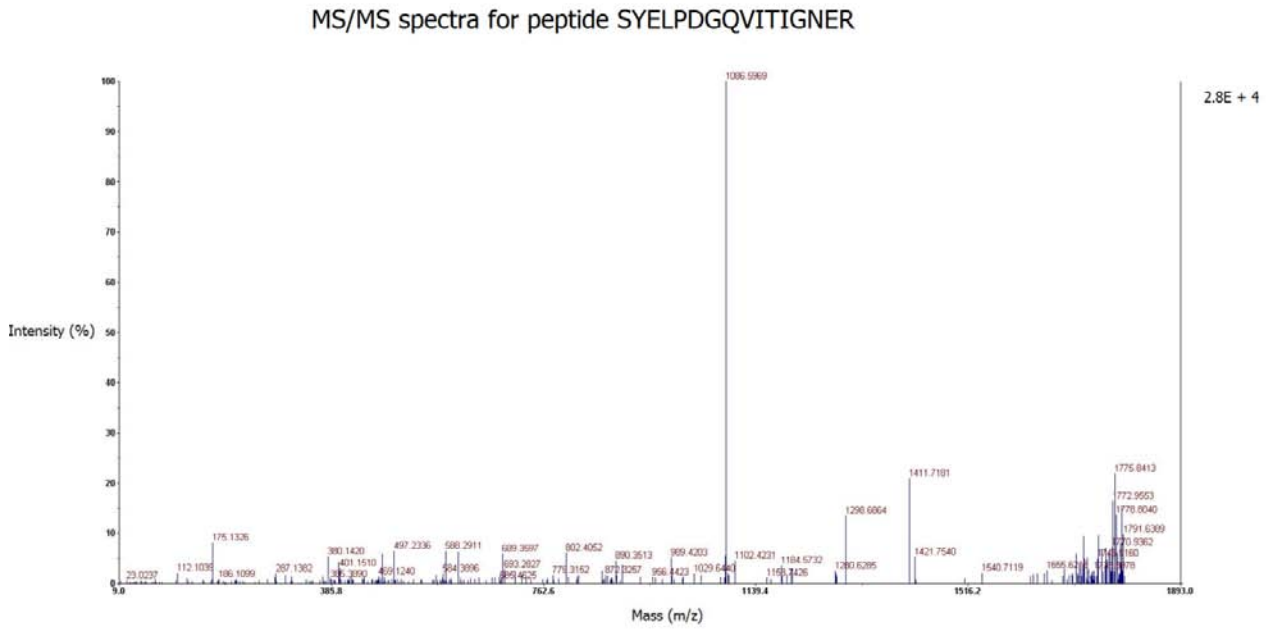
MS/MS spectrum annotated with observed masses, accompanied with a table of fragment ion assignments for peptide GHYTEGAELVDSVLDVVR. Since CID was used for MS/MS, only b- and y- series ions were shown. This peptide identified protein spot A9 as beta-tubulin.



b-series ions	Sequence	y-series ions
58.0287	**G**	
195.0877	**H**	1901.96
358.151	**Y**	1764.901
459.1987	**T**	1601.838
588.2413	**E**	1500.79
645.2627	**G**	***1371.748***
716.2998	**A**	1314.726
845.3424	**E**	1243.689
958.4265	**L**	***1114.647***
1057.495	**V**	***1001.563***
1172.522	**D**	***902.4942***
1259.554	**S**	***787.4672***
1358.622	**V**	700.4352
1471.706	**L**	***601.3668***
1586.733	**D**	***488.2827***
1685.802	**V**	***373.2558***
1784.87	**V**	***274.1874***
	R	***175.119***

Figure 4

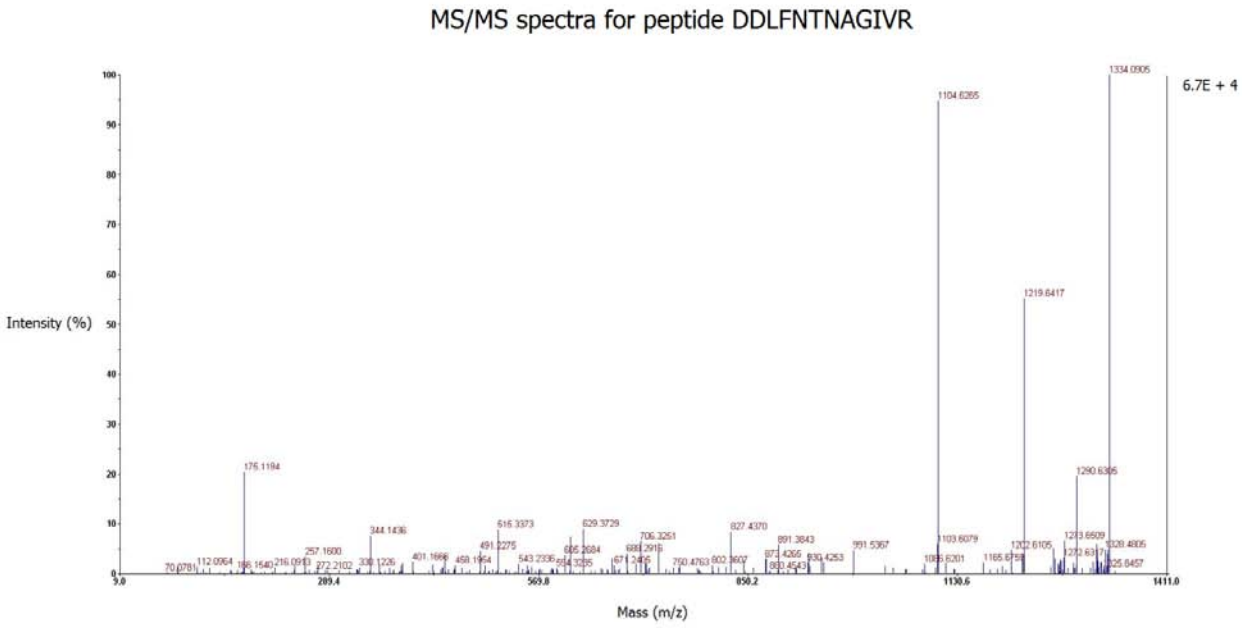
MS/MS spectrum annotated with observed masses, accompanied with a table of fragment ion assignments for peptide SYELPDGQVITIGNER. Since CID was used for MS/MS, only b- and y- series ions were shown. This peptide identified protein spot B11 as actin.



b-series ions	Sequence	y-series ions
88.0393	S	
251.1026	Y	1703.86
380.1452	E	1540.797
493.2293	L	***1411.754***
590.2821	P	***1298.67***
705.309	D	1201.617
762.3305	G	***1086.59***
890.389	Q	1029.569
989.4575	V	***901.5102***
1102.542	I	***802.4417***
1203.589	T	***689.3577***
1316.673	I	***588.31***
1373.695	G	***475.2259***
1487.738	N	418.2045
1616.78	E	***304.1615***
	R	***175.119***

Figure 5

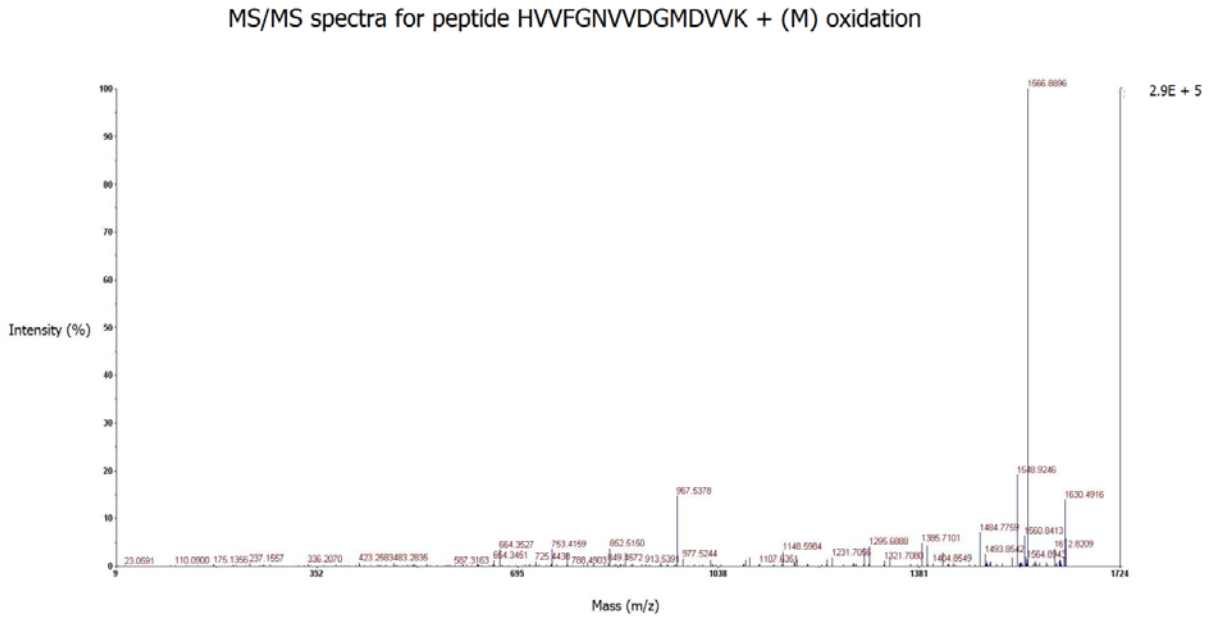
MS/MS spectrum annotated with observed masses, accompanied with a table of fragment ion assignments for peptide DDLFNTNAGIVR. Since CID was used for MS/MS, only b- and y-series ions were shown. This peptide identified protein spot D8 as malate dehydrogenase precursor



b-series ions	Sequence	y-series ions
116.0342	D	
231.0612	D	***1219.643***
344.1452	L	***1104.616***
491.2136	F	***991.532***
605.2566	N	***844.4635***
706.3042	T	***730.4206***
820.3472	N	***629.3729***
891.3843	A	***515.33***
948.4058	G	***444.2929***
1061.49	I	***387.2714***
1160.558	V	***274.1874***
	R	***175.119***

Figure 6

MS/MS spectrum annotated with observed masses, accompanied with a table of fragment ion assignments for peptide HVVFGNVVDGMDVVK + oxidation of methionine. Since CID was used for MS/MS, only b- and y- series ions were shown. This peptide identified protein spot G8 as peptidyl prolyl cis-trans isomerase A



b-series ions	Sequence	y-series ions
138.0662	H	
237.1346	V	1493.7668
336.203	V	***1394.6984***
483.2714	F	1295.63
540.2929	G	***1148.5616***
654.3358	N	***1091.5401***
753.4042	V	***977.4972***
852.4726	V	***878.4288***
967.4996	D	***779.3604***
1024.5211	G	***664.3334***
1171.5565	M	***607.312***
1286.5834	D	460.2766
1385.6518	V	345.2496
1484.7202	V	246.1812
	K	147.1128

Table 5. MS/MS ions search against the Invertebrate EST database

Spot ID[a]	Accession no.	Blast description	Individual ions score*	Expect	Error (ppm)[e]	m/z[c]	Sequence confirmed by MS/MS[d]
A11	EX000247	Beta-tubulin	81	0.00054	4.60	1130.6005	R.FPGQLNADLR.K
			105	9.4e-07	18.3	1959.0177	K.GHYTEGAELVDSVLDVVR.K
			73	0.0011	17.6	2087.1135	K.GHYTEGAELVDSVLDVVRK.E
C6	FY000758	Actin	97	8.2e-06	7.54	1790.9054	K.SYELPDGQVITIGNER.F
E18	DW263219	ATP synthase	187	3.7e-15	39.0	2408.2778	R.EVAAFAQFGSDLDQATQNLLNR.G
A5	FL489343	Tropomyosin	80	0.00039	0.78	1670.8357	R.TIDTHEQEIQSLTR.K
B16	FC567155	Gelsolin	120	3.5e-08	18.3	1869.9683	K.TVELDTFLDDAPIQHR.E
B20	ES394536	T-complex protein 1 beta	59	0.024	15.9	2292.2170	K.ILTQYKDHFSNLCVDAVLR.L
B21	ES394173	Tektin	92	1.6e-05	35.9	1614.9099	K.NLPTDVAIECLTLR.E
G6	FL488962	Enkurin	89	0.000035	12.9	1722.94	K.NLLEPSGLEPVYVHR.K

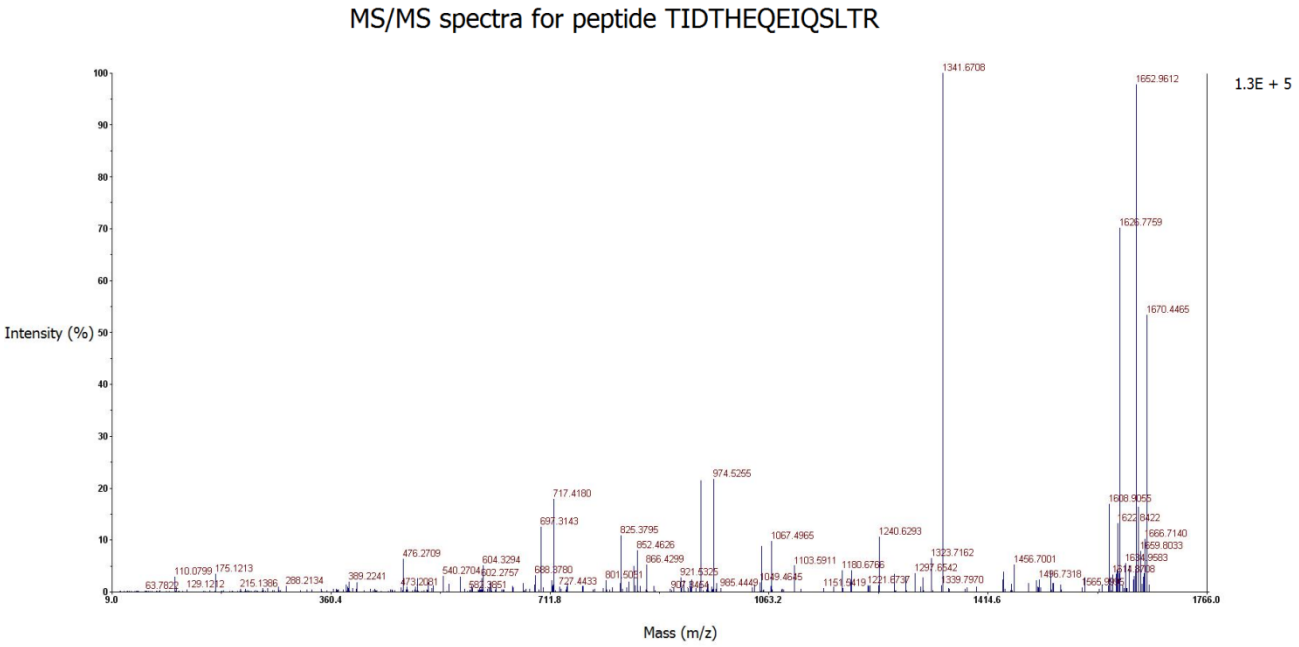
[a] refers to protein spots from figure 1a and 1b
[c] all peptides have a +1 charge
[d] sequences were confirmed by collision-induced dissociation
[e] parts per million (PPM)
[f] protein blast search result using the expressed sequence tag

* Individual ions scores greater than 57 were significant ($p<0.05$).

MS/MS ions search was carried out using Mascot (http://www.matrixscience.com/). Individual ion scores are based on the equation -10*Log(P), where (P) is the probability that the observed match is a random event.

Figure 7

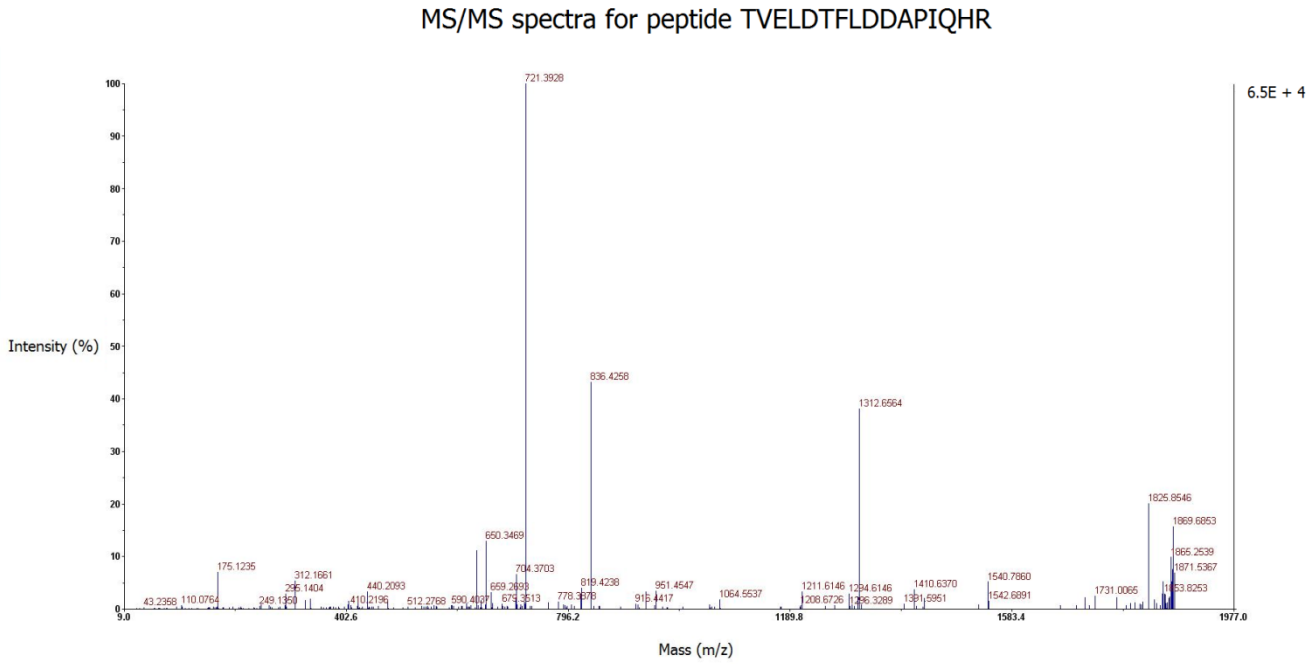
MS/MS spectrum annotated with observed masses, accompanied with a table of fragment ion assignments for peptide TIDTHEQEIQSLTR. Since CID was used, only b- and y- series ions were displayed.



b-series ions	Sequence	y-series ions
102.055	**T**	
215.139	**I**	1569.787
330.166	**D**	***1456.7***
431.2136	**T**	***1341.68***
568.273	**H**	***1240.63***
697.315	**E**	1103.569
825.374	**Q**	***974.527***
954.416	**E**	***846.468***
1067.5	**I**	***717.425***
1195.559	**Q**	***604.341***
1282.591	**S**	***476.283***
1395.675	**L**	***389.251***
1496.723	**T**	***276.167***
	R	***175.119***

Figure 8

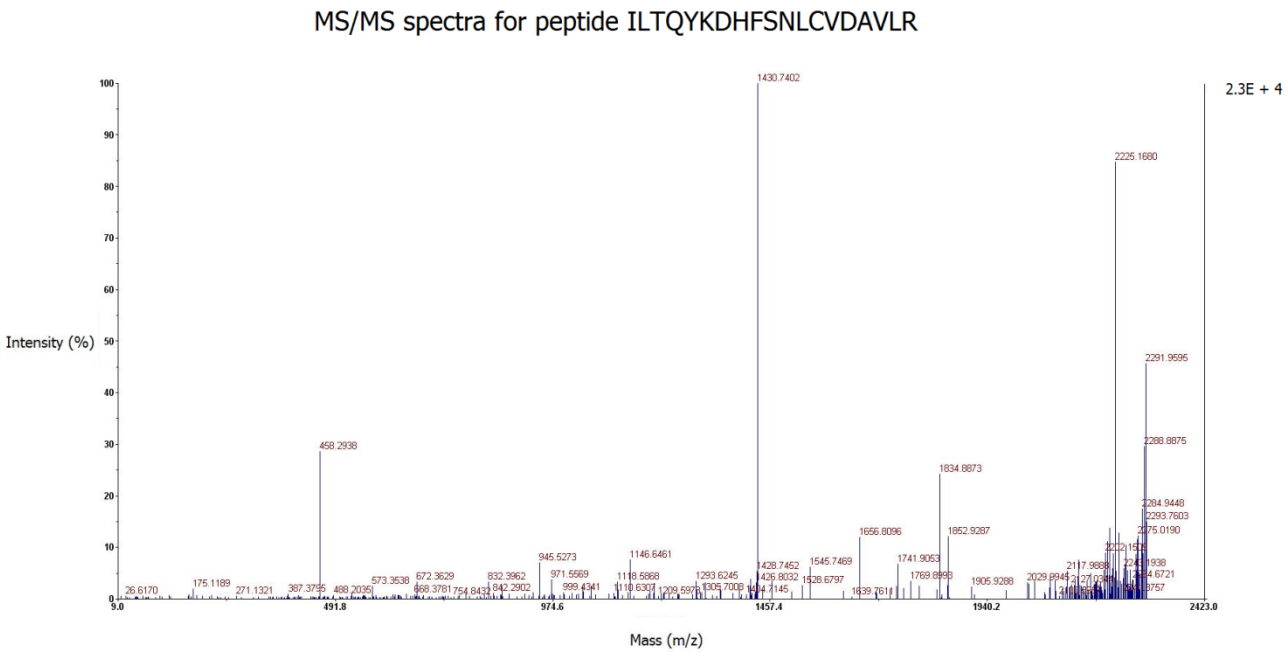
MS/MS spectrum annotated with observed masses, accompanied with a table of fragment ion assignments for peptide TVELDTFLDDAPIQHR. Since CID was used, only b- and y- series ions were displayed.



b-series ions	Sequence	y-series ions
102.055	**T**	
201.1234	**V**	***1768.89***
330.166	**E**	1669.818
443.25	**L**	***1540.78***
558.277	**D**	***1427.69***
659.3246	**T**	***1312.66***
806.3931	**F**	***1211.62***
919.4771	**L**	***1064.55***
1034.504	**D**	**951.4643**
1149.531	**D**	***836.437***
1220.568	**A**	***721.41***
1317.621	**P**	***650.373***
1430.705	**I**	553.3205
1558.764	**Q**	***440.236***
1695.822	**H**	***312.178***
	R	***175.119***

Figure 9

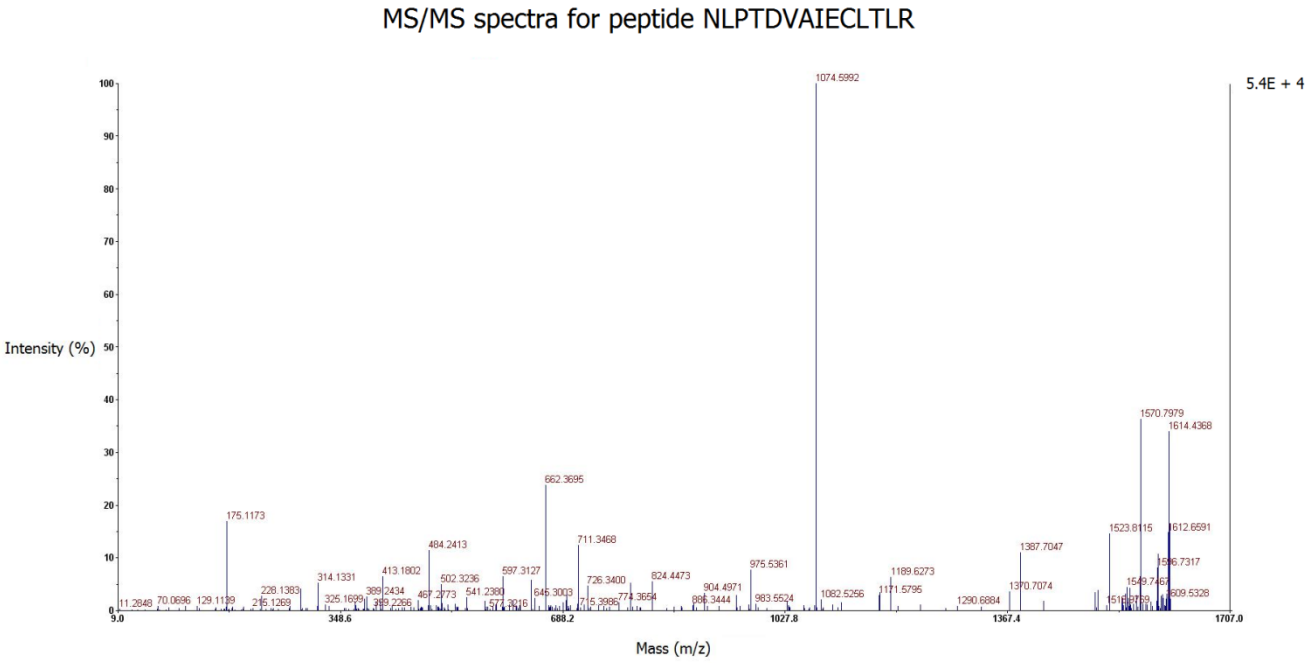
MS/MS spectrum annotated with observed masses, accompanied with a table of fragment ion assignments for peptide ILTQYKDHFSNLCVDAVLR. Since CID was used, only b- and y- series ions were displayed.



b-series ions	Sequence	y-series ions
114.0913	**I**	
227.1754	**L**	2179.097
328.2231	**T**	2066.012
456.2817	**Q**	1964.965
619.345	**Y**	1836.906
747.44	**K**	1673.843
862.4669	**D**	***1545.75***
999.5258	**H**	***1430.72***
1146.59	**F**	1293.662
1233.626	**S**	***1146.59***
1347.669	**N**	1059.562
1460.753	**L**	***945.519***
1620.784	**C**	***832.435***
1719.852	**V**	***672.404***
1834.88	**D**	***573.336***
1905.916	**A**	***458.309***
2004.985	**V**	387.2714
2118.069	**L**	288.203
	R	***175.119***

Figure 10

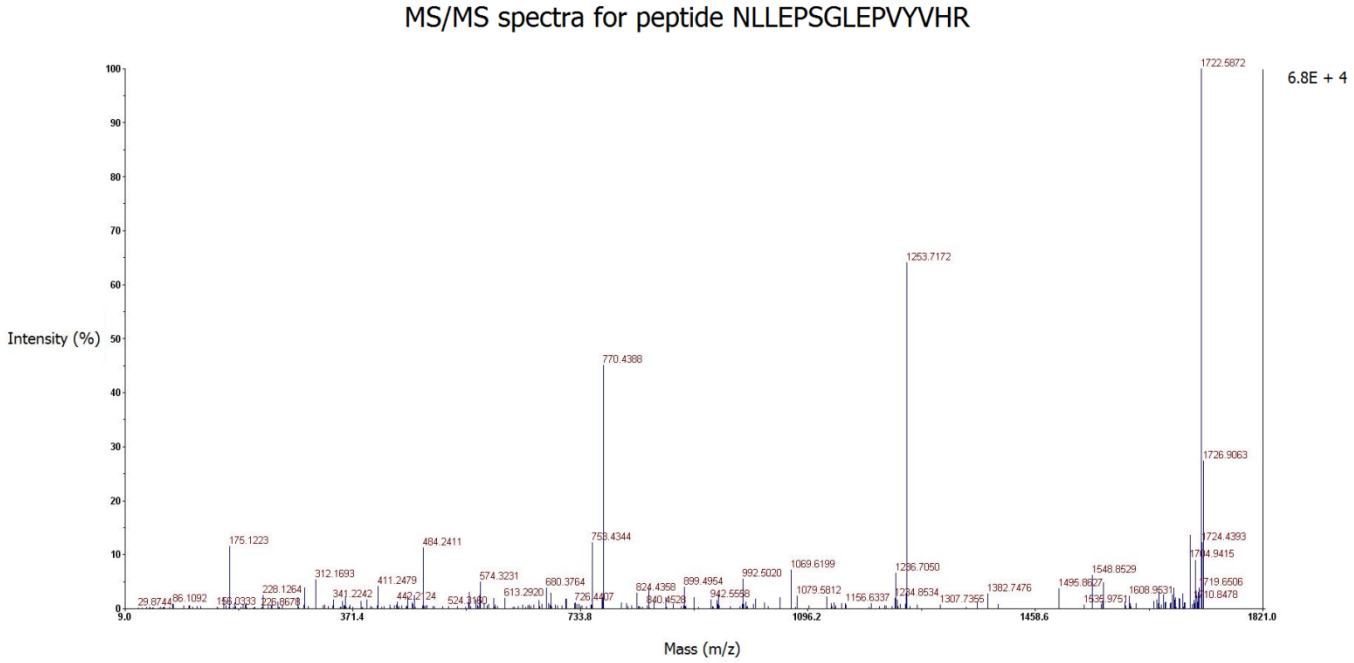
MS/MS spectrum annotated with observed masses, accompanied with a table of fragment ion assignments for peptide NLPTDVAIECLTLR. Since CID was used, only b- and y- series ions were displayed.



b-series ions	Sequence	y-series ions
115.0502	**N**	
228.134	**L**	1500.809
325.187	**P**	***1387.73***
426.2347	**T**	1290.672
541.2617	**D**	***1189.62***
640.33	**V**	***1074.6***
711.367	**A**	***975.529***
824.451	**I**	***904.492***
953.494	**E**	***791.408***
1113.525	**C**	***662.365***
1226.609	**L**	***502.335***
1327.656	**T**	***389.251***
1440.74	**L**	***288.203***
	R	***175.119***

Figure 11

MS/MS spectrum annotated with observed masses, accompanied with a table of fragment ion assignments for peptide NLLEPSGLEPVYVHR. Since CID was used, only b- and y- series ions were displayed.



b-series ions	Sequence	y-series ions
72.0444	**A**	
200.103	**Q**	1812.83
329.146	**E**	***1684.77***
458.188	**E**	***1555.73***
621.2515	**Y**	***1426.69***
736.2784	**D**	***1263.62***
807.316	**A**	***1148.6***
970.3789	**Y**	***1077.56***
1069.447	**V**	***914.496***
1140.484	**A**	***815.427***
1269.527	**E**	***744.39***
1406.586	**H**	***615.347***
1553.654	**F**	***478.289***
1709.76	**R**	***331.22***
	R	***175.119***

3.3 PEAKS DB

De novo sequences were searched against both the Mollusca and NCBInr database using PEAKS DB. This search strategy identified much of the same proteins as the Mascot database search (Table 5), except for protein spots A11 which was identified as ATP synthase beta subunit and F5 as heat shock protein 40. PEAKS DB also failed to identify protein spots A15, A22, B12, C9, D18 and F5-F19, which were originally identified by Mascot. PEAKS DB did manage to make two new protein identifications (Figure 7 and 8). Complement component C3-like protein (D12) was identified using a single *de novo* sequence, producing a probability-based score of 45.81, whereas phosphoenolpyruvate carboxykinase (F14) was identified from two *de novo* sequences with a score of 159.96.

Table 6. Summary results table for peptide identifications used for protein inference using PEAKS DB

Spot ID[a]	Accession no.[n]	Mr (kDa)	No: of peptides (coverage %)	Peptide sequence	-10lgP*	m/z[c]	Protein description
A1	gi|237862666	1024.4	2(13)	R.KPDGVFIINLR.K	139.55	1271.7512	Ribosomal protein SA
				R.FTPGTFTNQIQAAFREPR.L	93.74	2081.0764	
A2	gi|42559692	32.7	2(12)	K.QIQEHEQEIQSLTR.K	200	1738.8989	Tropomyosin
				K.NIQTENDYDNCNTQLQDVQAK.Y	113.86	2511.1921	
A9	gi|1174604	38.2	1(5)	K.GHYTEGAELVDSVLDVVR.K	126.26	1958.954	Beta-tubulin
A11	gi|51860821	49.8	9(18)	R.FPGQLNADLR.K	200	1130.6005	Beta-tubulin
				K.GHYTEGAELVDSVLDVVR.K	133.89	1959.0177	
				K.GHYTEGAELVDSVLDVVRK.E	111.37	2087.1135	
				R.YLTVAAM(+15.99)FR.G	49.04	1087.5609	
				R.LHFFM(+15.99)PGFAPLTSR.G	48.87	1636.849	
				K.IREEYPDR.I	41.3	1077.5328	
				K.LAVNM(+15.99)VPFPR.L	38.61	1159.6331	
				R.ISEQFTAM(+15.99)FR.R	20.22	1245.6011	
				R.KLAVNM(+15.99)VPFPR.L	18.14	1287.7328	
	gi|46909257	37.4	1(4)	K.AHGGYSVFAGVGER.T	200	1406.6975	ATP synthase beta subunit
A12	gi|53801335	4041.3	6(18)	R.FPGQLNADLR.K	200	1130.5845	Beta-tubulin
				K.GHYTEGAELVDSVLDVVRK.E	200	2087.0825	
				K.GHYTEGAELVDSVLDVVR.K	200	1958.9838	
				R.INVYYNEATGGKYVPR.A	133.11	1843.933	
				R.LHFFM(+15.99)PGFAPLTSR.G	112.27	1636.8241	
				K.LAVNM(+15.99)VPFPR.L	36.6	1159.6169	

A13	gi|1174593	50.2	7(22)	R.AVFVDLEPTVVDEVR.T	200	1687.8973	Tubulin alpha-2/alpha-4 chain
				R.QLFHPEQLITGKEDAANNYAR.G	200	2415.2202	
				R.IHFPLATYAPVISAEK.A	101.99	1756.9602	
				R.LIGQIVSSITASLR.F	93.05	1457.8674	
				R.QLFHPEQLITGK.E	70.34	1410.7616	
				K.VGINYQPPTVVPGGDLAK.V	69.05	1824.983	
				R.NLDIERPTYTNLNR.L	49.18	1718.8839	
A19	gi|46359618	73.1	3(6)	K.FDLTGIPPAPR.G	116.24	1183.6555	78 kDa glucose regulated protein
				R.IINEPTAAAIAYGLDKK.E	83.31	1788.0037	
				R.ARFEELNM(+15.99)DLFR.S	18.59	1556.7709	
A20	gi|46359618	73.1	2(4)	R.IINEPTAAAIAYGLDKK.E	127.53	1788.0261	78 kDa glucose regulated protein
				R.ARFEELNM(+15.99)DLFR.S	38.39	1556.7887	
A21	gi|153793258	83.1	1(1)	R.ALLFVPR.R	85.89	815.5519	Heat shock protein 90
A24	gi|224305	41.6	3(10)	K.SYELPDGQVITIGNER.F	200	1790.9609	Actin
				R.AVFPSIVGRPR.H	134.1	1198.7535	
				K.AGFAGDDAPR.A	112.64	976.4874	
B1	gi|89255272	41.1	5(18)	K.SYELPDGQVITIGNER.F	200	1790.8871	Cytoplasmic actin
				K.AGFAGDDAPR.A	118.82	976.4302	
				R.GYSFTTTAER.E	109.79	1132.5142	
				R.AVFPSIVGRPR.H	92.63	1198.6913	
				R.VAPEEHPVLLTEAPLNPK.A	89.6	1954.0602	
	gi|57635269	71.2	3(6)	K.STSGDTHLGGEDFDNR.M	148.88	1707.7119	Heat shock protein 70
				R.ARFEELNADLFR.G	95.98	1480.7443	
				K.DAGTISGM(+15.99)NVLR.I	20.55	1249.6299	
B2	gi|224305	41.6	4(15)	K.SYELPDGQVITIGNER.F	200	1790.9061	Actin
				R.VAPEEHPVLLTEAPLNPK.A	97	1954.0745	
				K.AGFAGDDAPR.A	90.97	976.4408	
				R.AVFPSIVGRPR.H	88.42	1198.7042	
	gi|147907866	66.0	2(4)	K.NAVVTVPAYFNDSQR.Q	200	1680.844	Mitochondrial mortalin splice variant
				K.DAGQISGLNVLR.V	97	1242.6801	
	gi|77023195	17.6	2(18)	K.STSGDTHLGGEDFDNR.M	90.97	1707.7356	Cytosolic heat shock cognate protein 70
				R.ARFEELNADLFR.G	88.42	1480.7587	

B3	gi|224305	41.6	4(13)	K.SYELPDGQVITIGNER.F	132.37	1790.9207	Actin
				K.AGFAGDDAPR.A	74.84	976.4498	
				R.AVFPSIVGRPR.H	55.5	1198.7164	
				R.GYSFTTTAER.E	36.24	1132.5394	
B4	gi|89255272	41.1	4(15)	K.SYELPDGQVITIGNER.F	200	1790.943	Cytoplasmic actin
				K.AGFAGDDAPR.A	200	976.4658	
				R.VAPEEHPVLLTEAPLNPK.A	127.62	1954.1161	
				R.AVFPSIVGRPR.H	69.77	1198.7343	
B5	gi|166406898	41.7	3(10)	K.SYELPDGQVITIGNER.F	200	1790.9009	Beta-actin 2
				R.AVFPSIVGRPR.H	119.08	1198.7051	
				K.AGFAGDDAPR.A	78.44	976.4466	
B6	gi|89255272	41.1	8(26)	K.SYELPDGQVITIGNER.F	200	1790.8943	Cytoplasmic actin
				K.AGFAGDDAPR.A	133.06	976.4357	
				R.VAPEEHPVLLTEAPLNPK.A	122.82	1954.0646	
				R.GYSFTTTAER.E	120.99	1132.5203	
				R.AVFPSIVGRPR.H	90.51	1198.6974	
				R.KDLYANTVLSGGSTM(+15.99)YPGIADR.M	67.7	2345.1526	
				K.IKIIAPPER.K	64.31	1036.6392	
				K.IIAPPERK.Y	61.27	923.5544	
B7	gi|159507454	41.8	5(23)	K.SYELPDGQVITIGNER.F	200	1790.9122	Beta-actin
				K.AGFAGDDAPR.A	144.71	976.4434	
				R.VAPEEHPVLLTEAPLNPK.A	117.08	1954.0863	
				R.AVFPSIVGRPR.H	115.87	1198.7085	
				R.GYSFTTTAER.E	88.2	1132.5304	
B8	gi|159507454	41.8	6(23)	K.SYELPDGQVITIGNER.F	200	1790.8879	Beta-actin
				R.GYSFTTTAER.E	132.41	1132.5198	
				K.AGFAGDDAPR.A	130.06	976.4349	
				R.VAPEEHPVLLTEAPLNPK.A	126.64	1954.0594	
				R.AVFPSIVGRPR.H	106.86	1198.6964	
				R.KDLYANTVLSGGSTM(+15.99)YPGIADR.M	51.93	2345.1506	

B10	gi|1174593	50.2	2(8)	R.AVFVDLEPTVVDEVR.T	200	1687.9216	Tubulin alpha-2/alpha-4 chain
				R.QLFHPEQLITGKEDAANNYAR.G	200	2415.2529	
B11	gi|116078087	14.8	1(12)	K.SYELPDGQVITIGNER.F	200	1790.8849	Actin
B13	gi|1335661	49.6	6(13)	K.GHYTEGAELVDSVLDVVR.K	200	1958.9896	Beta-tubulin
				K.GHYTEGAELVDSVLDVVRK.E	200	2087.0889	
				R.INVYYNEATGGKYVPR.A	200	1843.9386	
				R.INVYYNEATGGK.Y	200	1328.6338	
				R.AVLVDLEPGTM(+15.99)DSVR.S	136.39	1617.8113	
				K.IREEYPDR.I	83.33	1077.5182	
B15	gi|1174593	50.2	5(14)	R.AVFVDLEPTVVDEVR.T	200	1687.9127	Tubulin alpha-2/alpha-4 chain
				R.QLFHPEQLITGKEDAANNYAR.G	200	2415.2393	
				R.NLDIERPTYTNLNR.L	106.69	1718.9032	
				R.QLFHPEQLITGK.E	31.45	1410.7789	
				R.LIGQIVSSITASLR.F	5.89	1457.882	
C2	gi|1174593	50.2	5(18)	R.AVFVDLEPTVVDEVR.T	200	1687.9204	Tubulin alpha-2/alpha-4 chain
				R.QLFHPEQLITGKEDAANNYAR.G	200	2415.2583	
				R.LIGQIVSSITASLR.F	122.45	1457.8877	
				R.IHFPLATYAPVISAEK.A	98.95	1756.9862	
				R.NLDIERPTYTNLNR.L	61.85	1718.9098	
C3	gi|53801335	4041.3	9(25)	R.FPGQLNADLR.K	200	1130.5879	Beta-tubulin
				K.GHYTEGAELVDSVLDVVR.K	200	1958.9921	
				R.INVYYNEATGGKYVPR.A	114.74	1843.9381	
				R.LHFFM(+15.99)PGFAPLTSR.G	106.42	1636.8301	
				K.GHYTEGAELVDSVLDVVRK.E	86.91	2087.0889	
				K.IREEYPDR.I	62.97	1077.5211	
				R.ALTVPELTQQM(+15.99)FDAK.N	60.61	1707.8597	
				R.KLAVNM(+15.99)VPFPR.L	55.88	1287.7179	
				K.LAVNM(+15.99)VPFPR.L	42.67	1159.62	
C4	gi|116078087	14.8	1(12)	K.SYELPDGQVITIGNER.F	200	1790.9117	Actin

C5	gi|315572230	2018.1	2(21)	K.SYELPDGQVITIGNER	200	1790.8889	Actin
				R.VAPEEHPVLLTEAPLNPK.A	84.58	1954.0559	
	gi|121014	37.3	3(9)	R.ELPGHTGYLSC(+57.02)C(+57.02)R.F	121.18	1549.6874	Guanine nucleotide-binding protein subunit beta
				K.VHAIPLR.S	61.06	805.488	
				R.AGVLAGHDNR.V	22.64	1009.5085	
C6	gi|159507454	41.8	4(16)	K.SYELPDGQVITIGNER.F	200	1790.9054	Beta-actin
				R.GYSFTTTAER.E	200	1132.5359	
				K.QEYDESGPSIVHR.K	113.11	1516.7094	
				R.KDLYANTVLSGGSTM(+15.99)YPGIADR.M	59.58	2345.1572	
C8	gi|224305	41.6	2(8)	K.SYELPDGQVITIGNER.F	200	1790.9158	Actin
				K.QEYDESGPSIVHR.K	200	1516.7192	
C10	gi|1911573	47.4	1(1)	K.YNQILR.I	31.7	806.4549	Enolase
C12	gi|1911573	47.4	1(1)	K.YNQILR.I	61.76	806.4545	Enolase
C23	gi|58219310	1050.0	1(3)	R.AVFVDLEPTVVDEVR.T	141.76	1687.9701	Tubulin
D5	gi|209171293	19.0	1(8)	K.AYGVYLQDLGHSLR.G	200	1591.83	Peroxiredoxin 4 variant precursor
D8	gi|6746611	36.3	1(4)	R.DDLFNTNAGIVR.D	200	1334.6869	Malate dehydrogenase precursor
D13	gi|215263232	15.8	2(19)	R.TVVVHADIDDLGKGGHELSK.T	121.9	2090.1191	Superoxide dismutase
				R.LAC(+57.02)GVIGISK.V	115.05	1017.5862	
D19	gi|1174593	50.2	2(8)	R.QLFHPEQLITGKEDAANNYAR.G	200	2415.2942	Tubulin alpha-2/alpha-4 chain
				R.AVFVDLEPTVVDEVR.T	200	1687.9563	
D20	gi|62768593	21.9	2(15)	R.AVFPSIVGRPR.H	200	1198.7405	Actin A1b
				R.VAPEEHPVLLTEAPLNPK.A	112.11	1954.1276	
D21	gi|192383355	192.8	1(1)	K.LC(+57.02)YSYGLLALLKR.E	45.81	1569.8441	Complement component C3-like protein

E2	gi|164510076	8.5	4(57)	K.IQDKEGIPPDQQR.L	200	1523.792	Ubiquitin
				R.TLSDYNIQKESTLHLVLR.L	200	2130.1709	
				R.LIFAGKQLEDGR.T	122.86	1346.7521	
				K.ESTLHLVLR.L	53.16	1067.6187	
	gi|25991946	42.9	4(14)	K.IQDKEGIPPDQQR.L	200	1523.792	Poly-ubiquitin
				R.TLSDYNIQKESTLHLVLR.L	200	2130.1709	
				R.LIFAGKQLEDGR.T	122.86	1346.7521	
				K.ESTLHLVLR.L	53.16	1067.6187	
E9	gi|126697388	18.6	2(15)	R.GDFC(+57.02)IDVGR.N	200	1038.4637	Nucleoside diphosphate kinase B
				R.MMLGATNPLQSNPGTIR.G	78.28	1800.932	
E18	gi|116008297	59.7	2(6)	R.EVAAFAQFGSDLDQATQNLLNR.G	200	2408.2778	Mitochondrial H+ ATPase a subunit
				R.EAYPGDVFYLHSR.L	200	1553.8033	
F5	gi|256549334	35.7	1(4)	R.AVYDQFGEEGLK.N	140.07	1355.6449	Heat shock protein 40A
F14	gi|113207854	71.4	2(4)	R.FTC(+57.02)PASQC(+57.02)PIIHPK.W	119.21	1655.8046	Phosphoenolpyruvate carboxykinase
				R.TMYVIPFSM(+15.99)GPIGGPLSK.I	81.51	1910.9708	
G5	gi|126697474	29.0	8(31)	R.ELYSQC(+57.02)FDELIR.Q	200	1572.7462	Axonemal dynein light chain p33
				K.AAQQTDEILNSILPPR.E	200	1765.9568	
				R.LDVVNLQEELDR.R	115	1442.7563	
				R.LDVVNLQEELDRR.L	107.89	1598.8558	
				K.YDNPVLVSR.N	83.59	1062.5505	
				M(+15.99)IPPNASLVKYDNPVLVSR.N	66.66	2129.1653	
				K.KHTEEIQFLK.R	52.57	1272.6898	
				R.ETGIC(+57.02)PVRR.E	19.49	1087.5613	

G8	gi|289064181	17.5	1(9)	K.HVVFGNVVDGM(+15.99)DVVK.A	132.98	1630.8297	Peptidyl prolyl cis-trans isomerase A

[a] refers to protein spots from figure 1a and 1b
[c] all peptides have a +1 charge
[n] refers to identifications from searching the NCBInr database

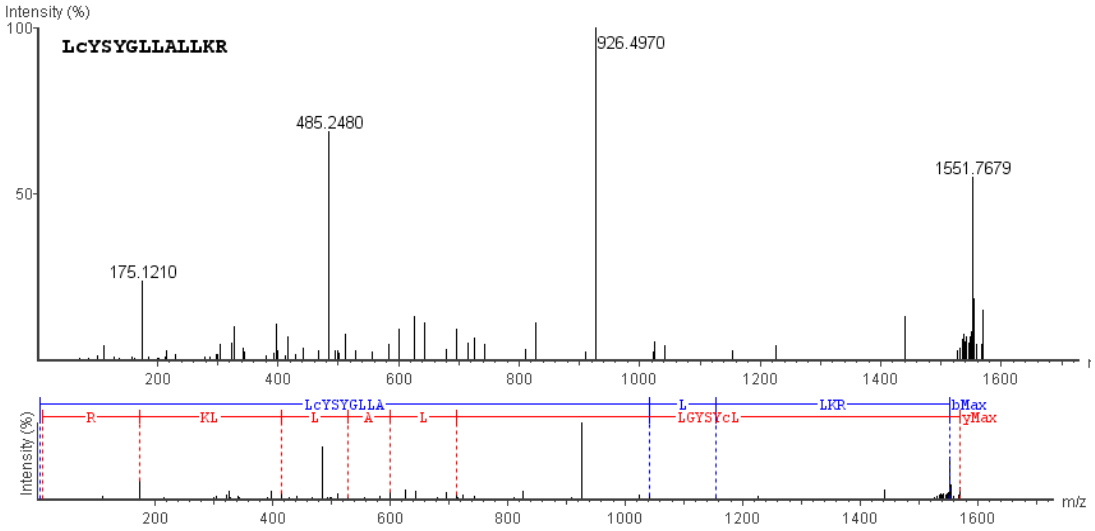
Included in this table is peptide identification confidence score, as well as matching peptide sequences and mass-to-charge ratio. The number of matching peptides and sequence coverage is also provided. Mr is also included to allow for the protein mass to be matched for new identifications.

* Peptide confidence score is given as -10lgP, where P is a false identification probability value. For small datasets (# spectra < 100) a -10lgP value of 20 has a 1% false positive identification rate.

The PEAKS DB search was carried out specifying the enzyme trypsin with a maximum of one missed cleavage. Carbamidomethylation of cysteine was selected as a fixed modification and oxidation of methionine as a variable modification. Monoisotopic mass values were selected along with a peptide mass tolerance of ± 50 ppm and fragment mass tolerance of ± 0.8 Da. A decoy search was also performed. Database searches were carried out against both the Mollusca and NCBInr protein sequence databases. The Mollusca database contained 58,900 sequences and NCBInr had 9,054,090 sequences.

Figure 12.

Annotated MS/MS spectrum accompanied with a spectrum alignment of the peptide LCYSYGLLALLKR for protein spot D21. Also included is a sequence alignment between the *de novo* and database sequence and peptide identification results obtained from a PEAKS DB search.

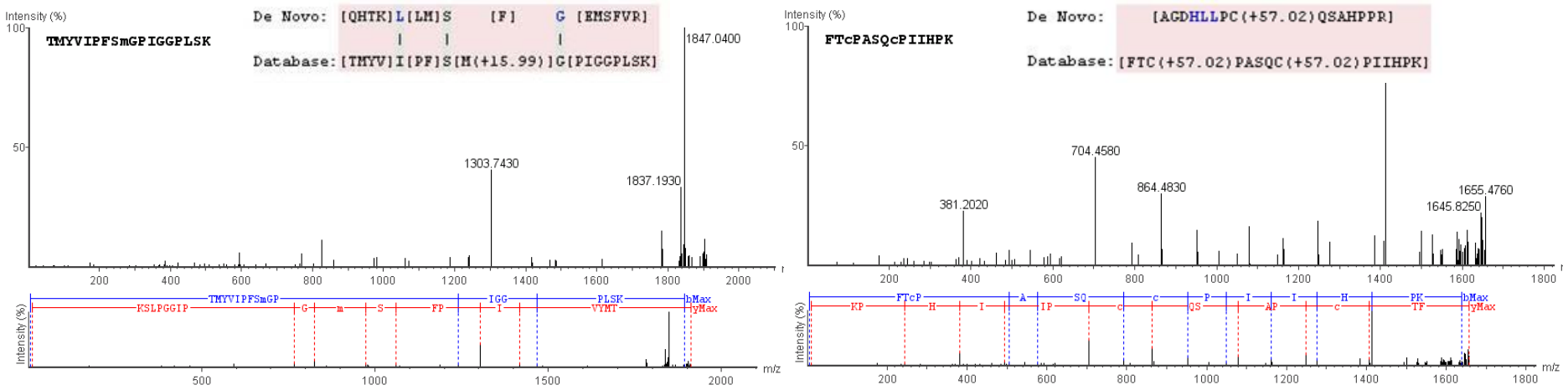



```
De Novo:     [ESQALDVLN]     [DYF] R
                                   |
Database: [LC(+57.02)YSYGLL] [ALLK] R
```

Accession no.	-10lgP	Mr (kDa)	Peptide sequence	Error (ppm)	m/z	Protein description
gi|192383355	45.81	192.8	LC(+57.02)YSYGLLALLKR	-24.3	1569.8441	Complement component C3-like protein

Figure 13.

Annotated MS/MS spectrum accompanied with spectrum alignments of the peptides FTCPASQCPIIHPK and TMYVIPFSMGPIGGPLSK for protein spot F14. Also included are sequence alignments between the *de novo* and database sequence and peptide identification results obtained from a PEAKS DB search.



Accession no.	-10lgP	Mr (kDa)	Peptide sequence	Error (ppm)	m/z	Protein description
gi|113207854	159.96	71.4	R.FTC(+57.02)PASQC(+57.02)PIIHPK.W(U)	0.8	1655.8046	Phosphoenolpyruvate carboxykinase
			R.TMYVIPFSM(+15.99)GPIGGPLSK.I(U)	-2.4	1910.9708	

3.4 SPIDER homology search

The SPIDER homology search involved searching *de novo* sequences against the Mollusca database only. This strategy managed to identify 50 proteins (Table 6), 10 of which were new identifications (Table 7). All of these proteins were identified from two or more *de novo* sequences, with cytosolic malate dehydrogenase (spot C19) producing the highest score of 116.08. Other proteins producing a confident SPIDER score were sodium-calcium exchanger (B14), Glutamate receptor (B18), Omega-crystallin (B22 and B23), CREB-binding protein (C21), Glutamate receptor (D6), Poly(A)-binding protein (D12), Arginine kinase (E21) and Glyceraldehyde 3-phosphate dehydrogenase (E23).

Table 7. Summary results table for a SPIDER homology search.

Spot ID[a]	Accession no.	Protein score*	Mr (kDa)	No. of peptides (% coverage)[b]	Protein description
A2	gi|42559692	69	32.7	2(5)	Tropomyosin
A11	gi|1174604	62.95	38.2	2(8)	Beta-tubulin
A12	gi|1174604	147.17	38.2	4(13)	Beta-tubulin
A13	gi|1174593	199.52	50.2	6(15)	Tubulin alpha-2/alpha-4 chain
A15	gi|218683627	125.47	60.9	3(11)	Heat shock protein 60
A19	gi|3023914	171.7	73.7	7(10)	78 kDa glucose-regulated protein
A20	gi|38683403	61.54	71.3	2(3)	Heat shock protein 70
A22	gi|148717303	81.38	91.6	3(4)	Glucose-regulated protein 94
A24	gi|56693681	55.37	41.8	2(6)	Actin ovestestis isoform
B1	gi|42494887	119.18	71.8	3(7)	Heat shock protein 70
	gi|315572292	79.5	19.7	2(15)	Actin
B2	gi|38683403	118.54	71.3	3(5)	Heat shock protein 70
	gi|47117881	55.66	41.9	2(6)	Actin
B3	gi|315572292	49	19.7	1(9)	Actin
B4	gi|315572293	49	19.7	1(9)	Actin
B5	gi|47117881	59.23	41.9	2(6)	Actin
B6	gi|56693681	125.12	41.8	5(11)	Actin ovestestis isoform
B7	gi|47117881	87.09	41.9	3(9)	Actin
B8	gi|47117881	84.75	41.9	3(9)	Actin
B10	gi|1174593	71.4	50.2	2(8)	Tubulin alpha-2/alpha-4 chain
B11	gi|289064185	37.05	22.6	2(11)	Peptidyl prolyl cis-trans isomerase B
B13	gi|30088884	135.8	50.3	4(7)	Beta tubulin
B14	gi|220683564	33.5	14.7	2(10)	Sodium-calcium exchanger
B15	gi|1174593	115.8	50.2	4(14)	Tubulin alpha-2/alpha-4 chain
B18	gi|29823896	35.74	98.7	2(1)	Glutamate receptor
B22	gi|399302	41.34	56.1	2(4)	Omega-crystallin
B23	gi|399302	41.51	56.1	2(4)	Omega-crystallin
C2	gi|1174593	161.79	50.2	4(11)	Tubulin alpha-2/alpha-4 chain
C3	gi|1174604	118.71	38.2	4(15)	Beta-tubulin
C4	gi|315572292	49	19.7	1(9)	Actin
C5	gi|1730218	62.67	37.3	2(6)	Guanine nucleotide-binding protein subunit beta
	gi|315572292	47.13	19.7	1(9)	Actin
C6	gi|315572292	79.5	19.7	2(15)	Actin

C8	gi|315572292	79.5	19.7	1(9)	Actin
C10	gi|3023702	65.66	47.4	2(8)	Enolase
C12	gi|3023702	172.25	47.4	5(18)	Enolase
C19	gi|73656362	116.08	36.4	3(11)	Cytosolic malate dehydrogenase
C21	gi|21307831	49.62	248.6	2(2)	CREB-binding protein
C23	gi|47117251	35.41	50.0	1(3)	Tubulin alpha-1 chain
D6	gi|29823896	35.44	98.7	2(1)	Glutamate receptor
D8	gi|6746611	89.71	36.3	2(11)	Malate dehydrogenase precursor
D12	gi|7689377	40.11	32.7	2(2)	Poly(A)-binding protein
D13	gi|255983837	93.37	15.9	4(33)	Superoxide dismutase
D19	gi|1174593	101.04	50.2	2(8)	Tubulin alpha-2/alpha-4 chain
D20	gi|47117881	61.36	41.9	2(6)	Actin
E2	gi|12240042	100.25	14.7	3(33)	Ubiquitin
E9	gi|124265190	119.5	16.9	4(24)	Nucleoside diphosphate kinase
E18	gi|116008297	142.31	59.7	4(13)	Mitochondrial H+ ATPase a subunit
E21	gi|296837083	59.03	39.3	2(10)	Arginine kinase
E23	gi|290463452	45.38	36.1	2(7)	Glyceraldehyde 3-phosphate dehydrogenase
F5	gi|256549334	64.39	35.7	2(8)	Heat shock protein 40A
F14	gi|113207854	48.31	71.4	2(5)	Phosphoenolpyruvate carboxykinase
G5	gi|126697474	184.87	29.0	6(24)	Axonemal dynein light chain p33
G8	gi|295824573	118.53	17.3	5(32)	Cyclophilin A
	gi|289064181	104.6	17.5	4(31)	Peptidyl prolyl cis-trans isomerase A

[a] refers to protein spots from figure 1a and 1b

[b] (% coverage) is calculated by diving the number of amino acids by the protein amino acid length

* Confident SPIDER assignments have a protein score of 30 or greater.

A SPIDER homology search was carried out by searching *de novo* sequences with an ALC value greater than 50% against the Mollusca database. *De novo* sequencing errors were taken into account by specifying leucine = isoleucine and lysine = glutamine. Carbamidomethylation of cysteine was selected as a fixed modification and oxidation of methionine as a variable modification. Fragment ion tolerance was also set to 0.8 Da. This table contains a confidence score for protein as well as the number of peptides and sequence coverage. Mr is also included to allow for the protein mass to be matched for new identifications.

Table 8 Peptide results for new protein identifications from a SPIDER homology protein reconstruction search.

Spot ID[a]	Accession no.	Protein score*	Peptide score	RSD[f]	Error (ppm)[e]	m/z[c]	Peptide sequence	Protein description
B14	gi|220683564	33.50	18	0.33	1.2	1397.7078	T.DLSFTLGNDFLR.E	Sodium-calcium exchanger
			15.5	0.4	3.6	1683.8762	K.TGKDLSFTLGNDLFR.E	
B18	gi|29823896	35.74	20.95	0.33	78.1	982.572	D.TVPEGNTHK.Q	Glutamate receptor
			14.78	0.38	87	882.556	T.TPKGNTHK.Q	
B22	gi|399302	41.34	22	0.12	13.6	1070.5299	N.LYDEFVER.A	Omega-crystallin
			19.34	0.38	19.5	1517.7931	S.LPVSLGDFYSYTR.N	
B23	gi|399302	41.51	24.97	0.12	32.1	1070.5404	N.LYDEFVER.A	Omega-crystallin
			19.5	0.38	23.4	1517.8121	S.LVPLSGDFYSYTR.N	
C19	gi|73656362	116.08	41.11	0.25	3.6	2306.1858	T.TKPDHSYELVKGLSLNDFSR.E	Cytosolic malate dehydrogenase
			40.97	0.12	16.3	1778.9725	K.KYAPSLAPENFTALTR.L	
			34	0.13	16.3	1650.8756	K.YAPSLPAENFTALTR.L	
C21	gi|21307831	49.62	30.64	0.55	48.3	2770.5015	T.MGTSTYNATAGPLASSGSTATLLGSAVQR.M	CREB-binding protein
			18.98	0.55	41.4	1159.6407	A.TPPPVQMPGVH.T	
D6	gi|29823896	35.44	20.74	0.33	68.8	982.5629	D.TVPEGNTHK.Q	Glutamate receptor
			14.7	0.38	84	882.5533	T.TPKGNTHK.Q	

D12	gi|7689377	40.11	21	0.45	23.9	1298.6951	R.GFGFVTFRDPR.A	Poly(A)-binding protein
			19.11	0.5	46.7	1541.8369	K.GWGFGFVTFRDPR.A	
E21	gi|296837083	59.03	36.53	0.33	33.8	2139.1411	G.NGHGQHTESVGGVYVLSNKR.R	Arginine kinase
			22.5	0.44	41.2	1807.8799	R.SHDGYSFPPC(+57.02)LSVEGR.R	
E23	gi|290463452	45.38	28	0.38	-25	1897.8702	K.PLLTYTDEDVVSQDFR.G	Glyceraldehyde 3-phosphate dehydrogenase
			17.38	0.25	31.2	819.4727	K.VGLNGFGR.I	

[a] refers to protein spots from figure 1a and 1b
[c] all peptides have a +1 charge
[e] parts per million (PPM)
[f] (RSD) relative standard deviation

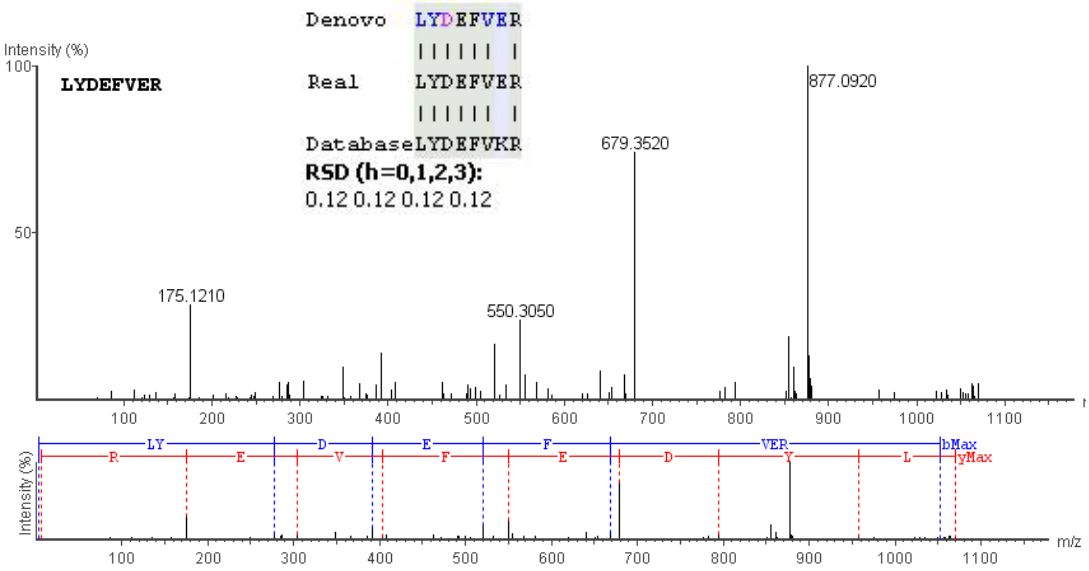
This table contains information about the peptide confidence score and RSD value. Peptide mass error is also included, along with the peptide sequence and mass-to-charge ratio.

* Confident SPIDER assignments have a protein score of 30 or greater and low RSD value (0.2 or lower, but can be higher in some cases)

The SPIDER homology search was carried out by searching *de novo* sequences with an ALC value greater than 50% against the Mollusca database*. De novo* sequencing errors were taken into account by specifying leucine = isoleucine and lysine = glutamine. Carbamidomethylation of cysteine was selected as a fixed modification and oxidation of methionine as a variable modification. Fragment ion tolerance was also set to 0.8 Da.

Figure 14.

Annotated MS/MS spectrum accompanied with a spectrum alignment of the peptides LYDEFVER and LVPLSGDFYSYTR for protein spot B23. Also included is a sequence alignment between the *de novo* and database sequence.



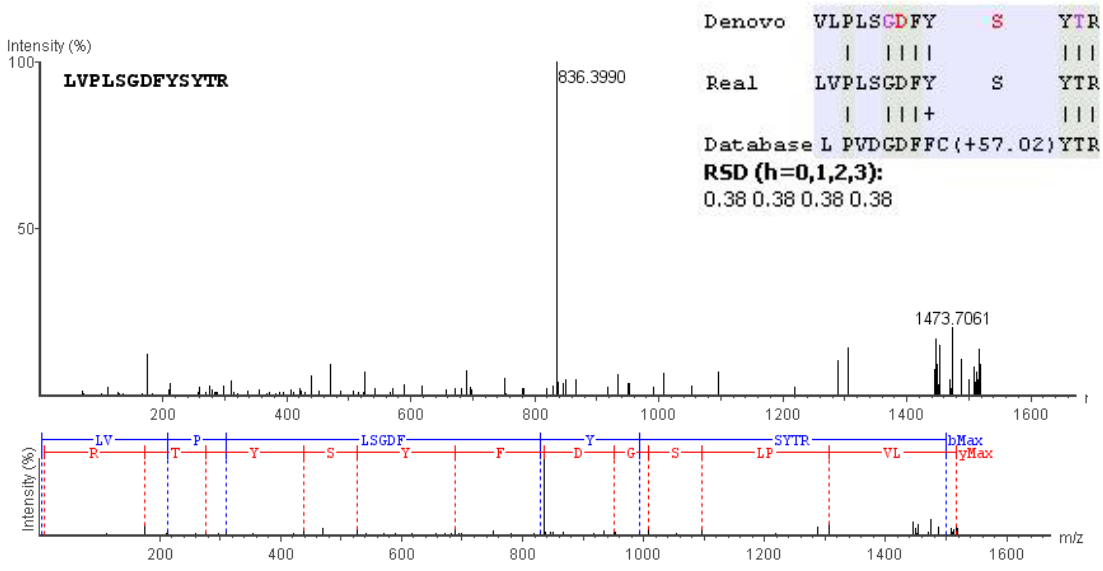


Figure 15.

Annotated MS/MS spectrum accompanied with a spectrum alignment of the peptides TKPDHSYELVKGLSLNDFSR and KYAPSLAPENFTALTR for protein spot C19. Also included is a sequence alignment between the *de novo* and database sequence.

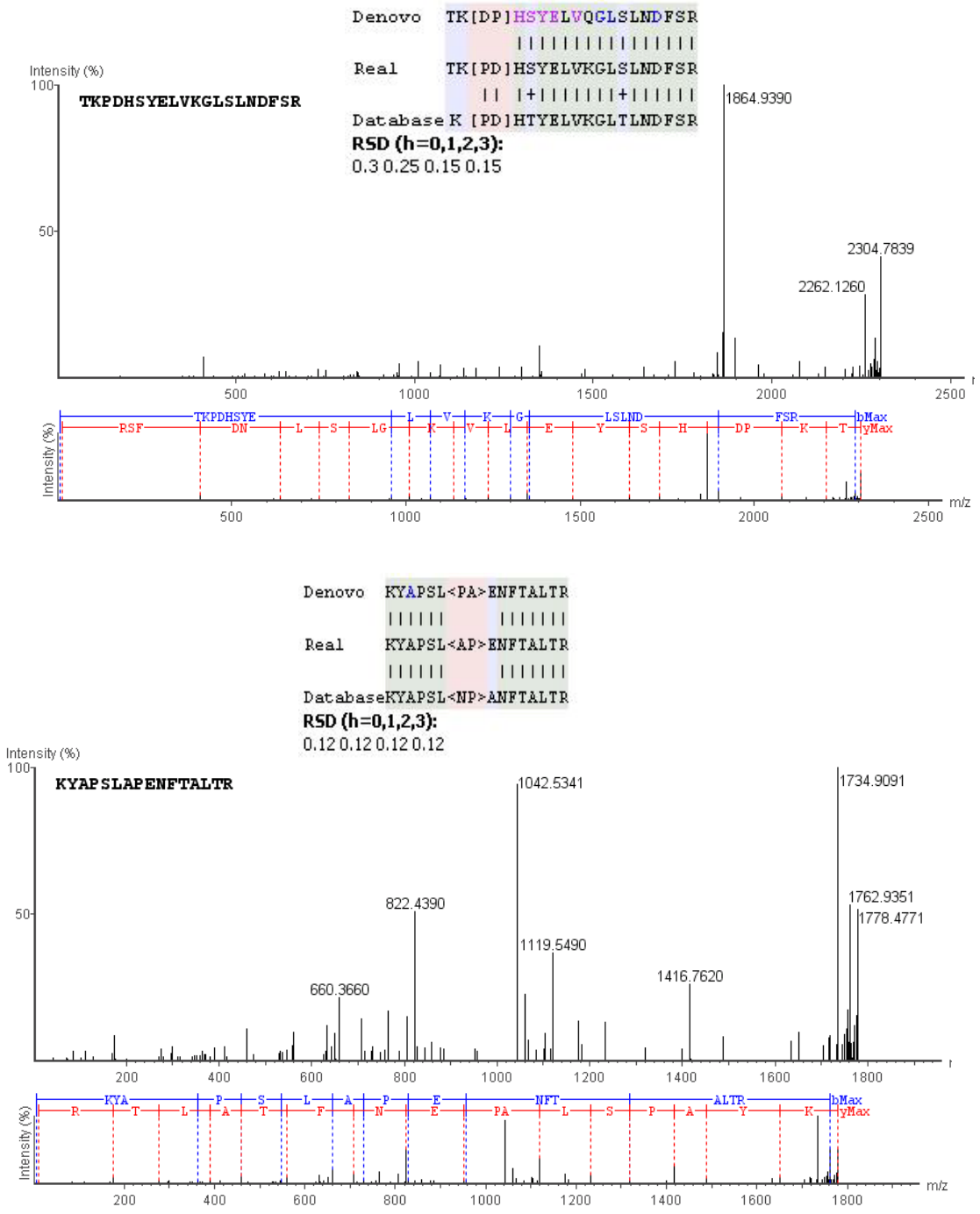


Figure 16.

Annotated MS/MS spectrum accompanied with a spectrum alignment of the peptides MGTSTYNATAGPLASSGSTATLLGSAVQR and TPPPVQMPGVH for protein spot C21. Also included is a sequence alignment between the *de novo* and database sequence.

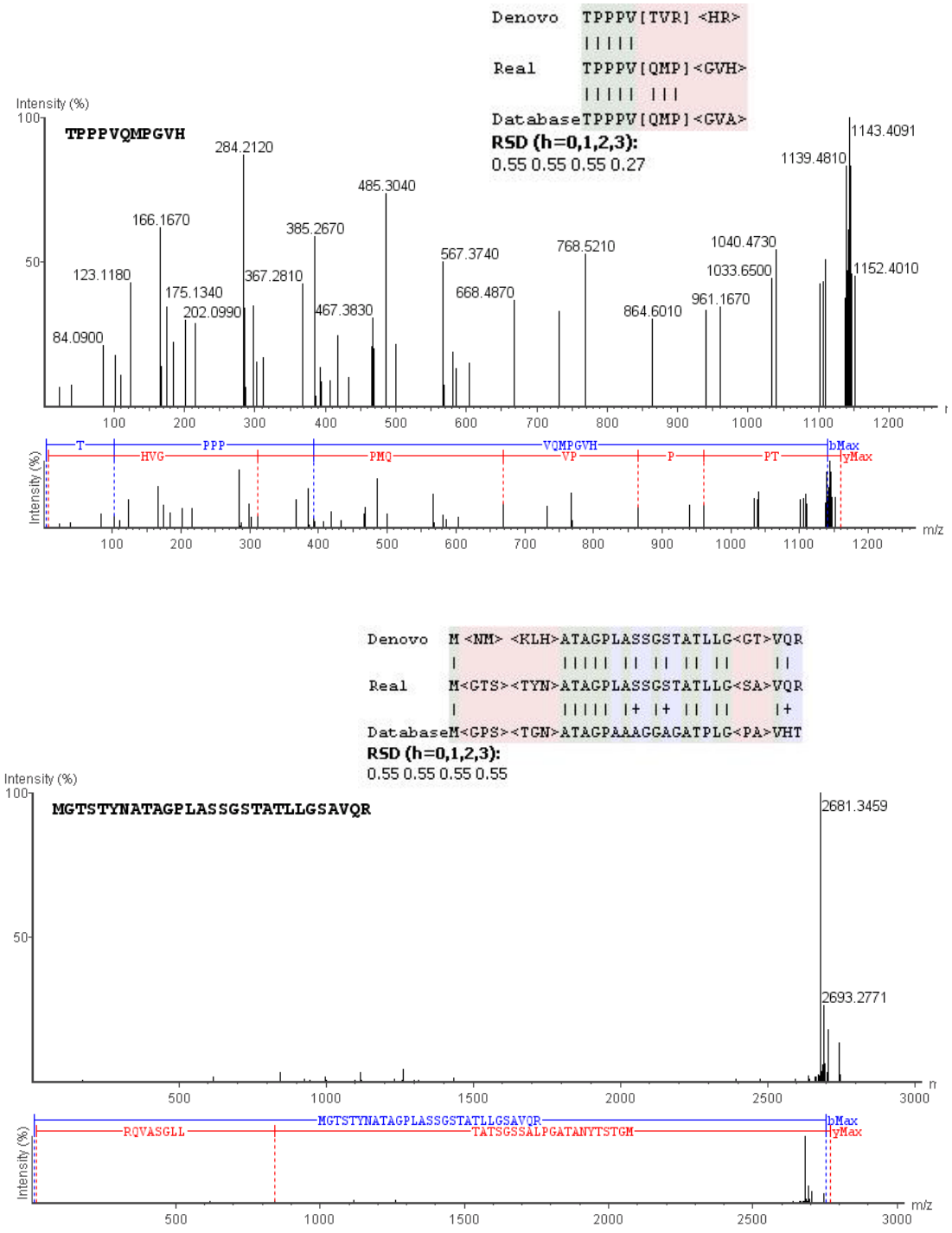
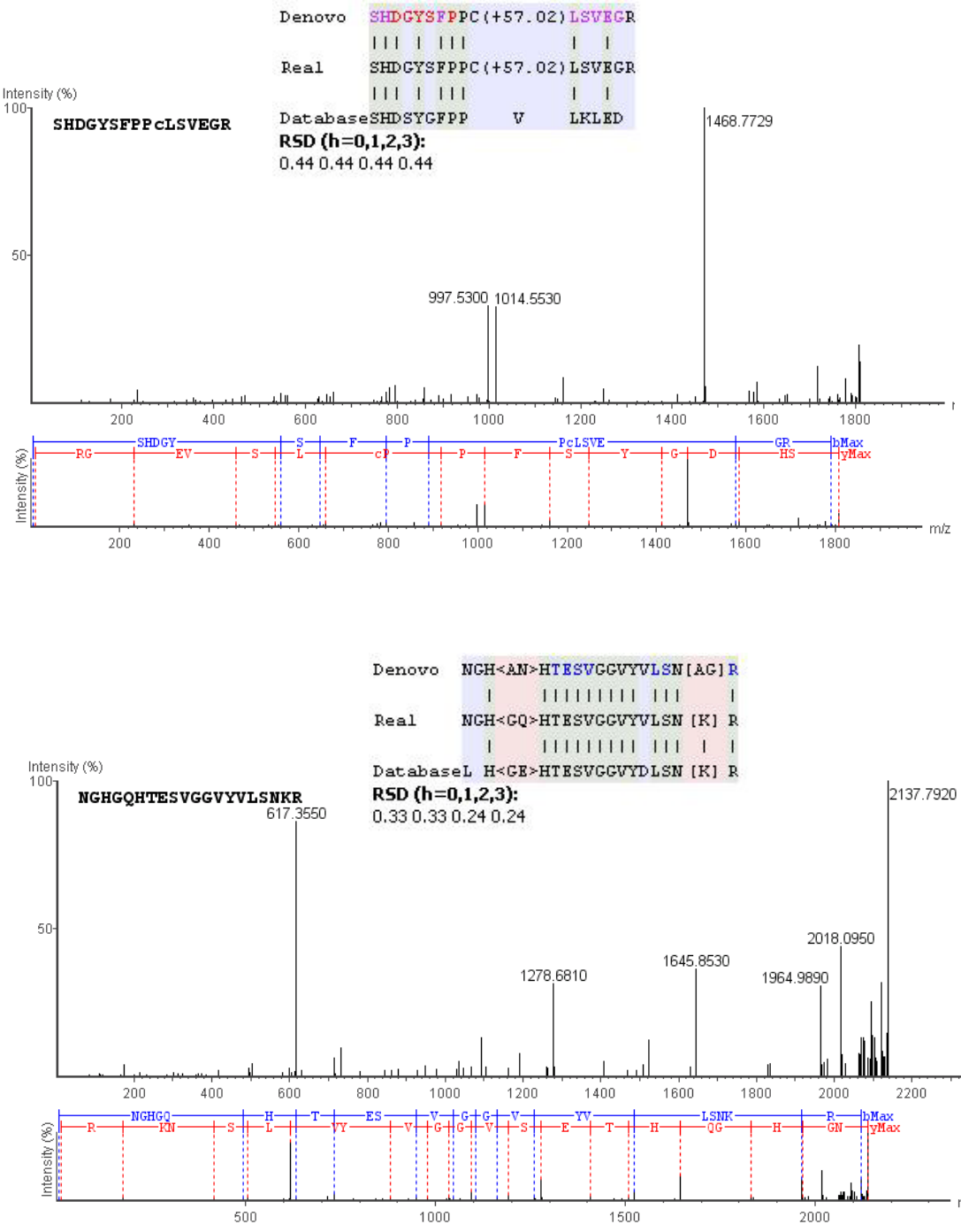


Figure 17.

Annotated MS/MS spectrum accompanied with a spectrum alignment of the peptides NGHGQHTESVGGVYVLSNKR and SHDGYSFPPCLSVEGR for protein spot E21. Also included is a sequence alignment between the *de novo* and database sequence.



3.5 Protein identification summary from all strategies

One-hundred and fifty-five protein spots of mid- to high-abundance were cut from the gels and analysed by mass spectrometry. In total, 73 proteins were identified: 57 by Mascot; an additional two using PEAKS DB; 10 from a SPIDER homology search and five by searching the Invertebrate EST database (Table 9). The majority of identifications were cytoskeletal proteins, with almost half identifying as actin or tubulin. Other noticeable groups include stress response proteins and those involved in protein biosynthesis. All remaining proteins had an assortment of functions.

Table 9. Summary results table for protein identifications from all strategies

Search strategy	Mascot						PEAKS DB		Spider	
Database	Mollusca		NCBInr		Invertebrate EST		Mollusca		Mollusca	
	Spot ID	Protein identity	SPOT ID	Protein identity	SPOT ID	Protein identity	Spot ID	Protein identity	Spot ID	Protein identity
	A1	40S ribosomal protein SA	A15	Heat shock protein 60	A5	Tropomyosin	D21	Complement C3-like protein	B14	Sodium-calcium exchanger
	A2	Tropomyosin	A22	Heat shock protein 90	B16	Gelsolin	F14	Phosphoenol-pyruvate Carboxykinase	B18	Glutamate receptor
	A9	Beta-tubulin	B12	Myosin light chain 2	B20	T-complex protein 1 beta			B22	Omega-crystallin
	A11	Beta-tubulin	C9	Eukaryotic translation factor 5A	B21	Tektin			B23	Omega-crystallin
	A12	Beta-tubulin	C10	Enolase 1	G6	Enkurin			C19	Cytosolic malate dehydrogenase
	A13	Tubulin alpha-2/alpha-4 chain	C12	Enolase 1					C21	CREB-binding protein
	A19	78 kDa glucose regulated protein	F5	L-threonine dehydrogenase					D6	Glutamate receptor
	A20	78 kDa glucose regulated protein	F9	Elongation factor 2					D12	Poly(A)-binding protein
	A21	Heat shock protein 90	F10	Elongation factor 2					E21	Arginine kinase
	A24	Actin	F11	Elongation factor 2					E23	Glyceraldehyde 3-phosphate dehydrogenase

	B1	Cytoplasmic actin	F18	Serine hydroxymethyl-transferase			
	B2	Actin	F19	GND1			
	B3	Actin					
	B4	Cytoplasmic actin					
	B5	Beta-actin					
	B6	Cytoplasmic actin					
	B7	Beta-actin					
	B8	Beta-actin					
	B10	Tubulin alpha-2/alpha-4 chain					
	B11	Actin					
	B13	Beta-tubulin					
	B15	Tubulin alpha-2/alpha-4 chain					
	C2	Tubulin alpha-2/alpha-4 chain					
	C3	Beta tubulin					
	C4	Actin					

	C5	Beta-actin				
	C6	Beta-actin				
	C8	Beta-actin				
	C23	Tubulin alpha-2/alpha-4 chain				
	D5	Peroxiredoxin 4 variant precursor				
	D8	Malate dehydrogenase precursor				
	D13	Superoxide dismutase				
	D18	Actin				
	D19	Tubulin alpha-2/alpha-4 chain				
	D20	Actin				
	E2	Polyubiquitin				
	E9	Nucleoside diphosphate kinase B				
	E18	Mitochondrial H+ ATPase alpha subunit				

	G5	Axonemal dynein light chain p33				
	G8	Peptidyl prolyl cis-trans isomerase A				
Total	39		12	5	2	10

Protein identifications were made by performing either a Mascot database search, PEAKS DB search or a Spider homology search. The Mascot search was carried out using a Mollusca, NCBInr or Invertebrate EST database, while both the PEAKS DB and Spider homology search was carried out using the Mollusca database only.

4. Discussion

4.1 Protein identifications

Almost all identifications were housekeeping proteins responsible for basic cellular functions. Housekeeping proteins face strong selective constraints and contain conserved regions that evolve more slowly than other proteins (Zhang and Li, 2004; She, Rohl *et al*., 2009). It is important for proteins to retain conserved regions for reasons including being part of a multi-subunit complex or interacting directly with other proteins (Krogan *et al*., 2006; Guharoy and Chakrabarti, 2010). Conserved regions also vary in size ranging from only a few amino acids to more than 60, with larger conserved regions easier to identify (Bejerano *et al*., 2004; Ren *et al*., 2008). Each of the protein groups identified - cytoskeletal proteins, stress response proteins, protein biosynthesis - are known to be highly conserved between species.

4.1.1 Cytoskeletal proteins

The cytoskeleton is a protein scaffold that is required to maintain cell shape, structure and function. It is made of microfilaments and microtubules, which are constructed from evolutionary conserved proteins (Wickstead and Gull, 2011). Microfilaments are made of actin, which exists as alpha, beta and gamma isoforms (Herman, 1993). The beta isoform was identified in the gill tissue of *Perna canaliculus* and is known to be found in non-muscle tissue (Ingerslev *et al*., 2006). Actin-binding proteins tropomyosin and gelsolin were also identified. Tropomyosin functions in intracellular transport, while gelsolin regulates the rearrangement of the cytoskeleton (Silacci *et al*., 2004). Tubulin and axonemal dynein also featured among identifications. Tubulin polymers make up microtubules, while axonemal dynein facilitates intracellular transport along microtubules in cilia (Fletcher and Mullins, 2010). Two other cilia-related proteins identified were tektin and enkurin

4.1.2 Stress response proteins

When cells encounter stress stimuli they launch a response by upregulating stress response proteins. Cellular stress can arise for several reasons, including exposure to heavy metals or oxidative stress (Farrer and Pecoraro, 2002; Kasprzak, 2002). Since stress response proteins operate as a cellular survival mechanism, they face strong selective constraints and are therefore highly conserved between species (Fulda *et al*., 2010). Proteins involved in the stress response include molecular chaperones (Kültz, 2003). The molecular chaperones identified include heat shock protein 60, 70 and 90, whose functions range from protein folding to inhibiting apoptosis (Jäättelä, 1999; Vargas-Parada *et al*., 2001). The 78 kDa glucose-regulated protein also belongs to the heat shock protein 70 family and can also protect cells against apoptosis (Luo *et al*., 2006).

Other proteins involved in the stress response also include antioxidant enzymes and those involved in protein biosynthesis. Antioxidant enzymes protect the cell from oxidative damage arising from free radicals, which is known to damage almost all cellular components (Dröge, 2002). Two antioxidant enzymes identified were superoxide dismutase and peroxiredoxin 4. Superoxide dismutase functions by converting superoxide into hydrogen peroxide (Deby and Goutier, 1990), while peroxiredoxin 4 protects cells by reacting with hydrogen peroxide (Tavender and Bulleid, 2010). Protein biosynthesis also plays a role in the stress response. Those identified included the highly conserved 40S ribosomal protein, translation initiation factor 5A and elongation factor 2, all of which are crucial for protein synthesis.

4.2 Evaluation of methods

4.2.1 Two-dimensional gel electrophoresis

2-DE was shown to be useful for separating proteins from the gill tissue of *Perna canaliculus*. More than 650 proteins were resolved along a pH gradient of 4-11, with the only major problem encountered being the over abundance of actin and tubulin. Not only did they obscure other proteins, but they also spilled over onto other proteins resulting in two incorrect identifications. Another problem was the poor detection of low abundance proteins. A strategy that can be used to overcome these issues is to first fractionate the sample prior to 2-DE. For example, fractionating using polyethylene glycol has been demonstrated to remove proteins of high abundance, as well as to improve the detection of low abundance proteins by up to five-fold (Xi *et al*., 2006).

4.2.2 Peptide analysis

Good MS spectra were obtained for most proteins, but not all. Despite all proteins being of high- to mid-abundance, at least 45 produced poor MS spectra. In some cases this was the result of the strong signal produced by interference spectra. It was difficult to pinpoint the exact source of the interference spectra, except when they arose from trypsin autolysis products or matrix particles. Interference spectra can arise from several sources, including chemical noise (Keller *et al*., 2008). Techniques that have been demonstrated to reduce chemical noise is to wash the MALDI plate with diammonium citrate after the peptides and matrix have co-crystalised or to add ammonium phosphate to the matrix (Smirnov *et al*., 2004).

Another possible reason for poor MS spectra is due to low peptide recovery. The amount of peptides recovered is heavily dependent on the peptide extraction method used. In another study, the use of acetonitrile was shown to be responsible for peptide losses of up to 50% (Speicher *et al*., 2000). A C18 Empore Disk has recently been demonstrated to improve peptide recovery and was found to be considerably more superior than ZipTips (Meng *et al*., 2008).

A third reason for poor MS spectra may be due to the matrix used. Although CHCA is regarded as the gold standard for peptide analysis, it is known to favour peptides with an arginine residue (Krause *et al*., 1999). This discriminatory feature may prevent some peptides from being analysed by mass spectrometry. Instead, a different matrix may be more useful. 4-Chloro-α-cyanocinnamic is a derivative of CHCA and has been reported to be less discriminative than its counterpart (Jaskolla *et al*., 2008). When both matrixes were compared, 4-chloro-α-cyanocinnamic achieved 44% greater sequence coverage than CHCA for an in-gel BSA digest.

4.3 Evaluation of search strategies

4.3.1 Mascot database search

Mascot was shown to be an effective search strategy for identifying proteins in *Perna canaliculus.* In total, 61 proteins were identified: 44 using the Mollusca database, 12 using an NCBInr database search and five by searching against the invertebrate EST database. Improvements to the Mollusca protein sequence database and the isolation of high abundance, highly conserved proteins were all major contributing factors to these identifications. An NCBInr database search was useful for identifying highly conserved proteins not found in the Mollusca database, while the invertebrate EST database search was useful for identifying proteins that could not be identified using any other search strategy.

4.3.2 PEAKS DB search

The purpose of a PEAKS DB search was to identify proteins that could not be identified by a Mascot database search. Despite 40 proteins producing two or more *de novo* sequences, only two new proteins were identified. These identifications included complement C3-like protein and phosphoenolpyruvate carboxykinase. Complement C3-like protein operates as part of the defence system, while phosphoenolpyruvate carboxykinase participates in cellular respiration and contains highly conserved histidine residues (Bazaes *et al*., 1997;

Venier *et al*., 2011). These conserved residues however were not responsible for its identification. Instead, the match was made using a leucine, serine and glycine residue. This demonstrates the power of a PEAKS DB search when comparable sequences are available in the database.

The most likely reason why these proteins were not identified is due to errors in the *de novo* sequence. The accuracy of *de novo* sequences can vary in the range of 18-49 % and is dependent on the quality of the MS/MS spectra (Pevtsov *et al*., 2006). But in many cases it is the result of amino acids sharing a similar mass. For instance, leucine and isoleucine have an equivalent mass, while only 0.036 Da separates lysine and glutamine (Ma and Johnson, 2012). A possible way around this is to use an error-tolerant search that takes into account *de novo* sequence errors.

4.3.3 SPIDER homology search

A homology search was carried out using SPIDER to identify proteins that could not be identified using PEAKS DB. SPIDER uses an error-tolerant search functionality to account for *de novo* sequencing errors and also allows for substitution, insertion and deletion mutations. This search strategy managed to identify 10 new proteins, each with a minimum of two *de novo* sequence matches. The only surprising identifications were for glutamate receptor and omega crystallin. Glutamate receptor is found in nerve tissue, whereas omega crystallin is found in the eye lens of scallops, squid and octopus (Dietz *et al*., 1992; Piatigorsky *et al*., 2000). However, gill tissue is known to be innervated, while omega crystallin is an inactive form of aldehyde dehydrogenase which can be found in several other species (Burleson and Smith, 2001; Horwitz *et al*., 2006). Other proteins identified were enzymes associated with cellular respiration, messenger RNA transport and a sodium-calcium exchanger.

4.4 Conclusion

The application of several different protein identification strategies were shown to be useful for identifying proteins in the greenshell mussel *Perna canaliculus*. This is an important finding since the inadequate representation of the greenshell mussel in sequence databases can be detrimental to future proteomic studies involving this species. The findings of this research can also be used to assist protein identification studies in other species poorly represented in sequence databases. On the whole, Mascot and PEAKS DB performed equally well, while the error-tolerant functionality of SPIDER was useful for identifying additional proteins. A search of the Invertebrate EST database was also useful for producing additional identifications. Although this workflow could be improved, it stands to reason improvements in the Mollusca database will inevitably result in more proteins being identified.

4.5 Future research

Future work should first focus on fractionating the protein sample to deplete high abundance proteins, while enriching low abundance proteins. This should improve the capacity of 2-DE to resolve more proteins which can then be identified using Mascot, PEAKS DB or SPIDER. It is also important that future work focuses on obtaining good quality MS spectra. This may be achieved by using a different peptide extraction technique or matrix, such as the C18 Empore Disk and 4-Chloro-α-cyanocinnamic acid. The option of using a different protease in place of trypsin, for example Lys-N, should also be considered. Since MS spectra are a precursor to obtaining MS/MS spectra, acquiring good quality MS spectra would also assist MS/MS-based identification strategies.

Database searches should also be extended to include genomic databases for the greenshell mussel *Perna canaliculus*. Molluscs contain considerably more genomic content in databases than protein sequences. Although there are issues surrounding false-positive identifications, a corresponding search against an EST database can help verify these results (Fermin *et al*., 2006). Mascot has the capabilities for carrying out a database search against a genomic database, while Indexed Genomes Gracefully Yield Peptide IDs or IggyPep can be used for searching *de nov*o sequences. IggyPep was recently shown to outperform Mascot in a search, identifying an additional 15 proteins (Menschaert *et al*., 2010). A database search could be further extended to include an RNA-seq database. Recent studies have reported an RNA-seq database to be useful for identifying novel peptides, even in species with a sequenced genome (Bitton *et al*., 2010; Wang *et al*., 2012).

Although there is a wealth of different search strategies that can be used, two potentially useful strategies are a MS-BLAST search and ByOnic. MS-BLAST is a sequence similarity search option that searches *de novo* sequences against a database. It has previously been successfully applied to several species with an unsequenced genome, identifying up to 70% of proteins (Grossmann *et al*., 2007; Ward *et al*., 2010). ByOnic uses a completely different approach and searches both peptide masses and flanking fragment ions against a database (Bern *et al*., 2007). ByOnic was capable of detecting low abundance peaks and was shown to be more sensitive than Mascot.

References

Aebersold, R. and M. Mann (2003). "Mass spectrometry-based proteomics." Nature 422(6928): 198-207.

Al-Subiai, S. N., A. J. Moody, et al. (2011). "A multiple biomarker approach to investigate the effects of copper on the marine bivalve mollusc, Mytilus edulis." Ecotoxicology and Environmental Safety 74(7): 1913-1920.

Baraj, B., F. Niencheski, et al. (2011). "Assessing the effects of Cu, Cd, and exposure period on metallothionein production in gills of the Brazilian brown mussel Perna perna by using factorial design." Environmental Monitoring and Assessment 179(1-4): 155-162.

Barboza, R., D. Cociorva, et al. (2011). "Can the false-discovery rate be misleading?" Proteomics 11(20): 4105-4108.

Bazaes, S., L. Montecinos, et al. (1997). "Identification of reactive conserved histidines in phosphoenolpyruvate carboxykinases from Escherichia coli and Saccharomyces cerevisiae." Biochimica et Biophysica Acta - Protein Structure and Molecular Enzymology 1337(2): 166-174.

Beavis, R. C., T. Chaudhary, et al. (1992). "α-Cyano-4-hydroxycinnamic acid as a matrix for matrixassisted laser desorption mass spectromtry." Organic Mass Spectrometry 27(2): 156-158.

Bejerano, G., M. Pheasant, et al. (2004). "Ultraconserved elements in the human genome." Science 304(5675): 1321-1325.

Bern, M., Y. Cai, et al. (2007). "Lookup peaks: A hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry." Analytical Chemistry 79(4): 1393-1400.

Bitton, D. A., D. L. Smith, et al. (2010). "An integrated mass-spectrometry pipeline identifies novel protein coding-regions in the human genome." PloS one 5(1).

Brosch, M., G. I. Saunders, et al. (2011). "Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome." Genome Research 21(5): 756-767.

Burleson, M. L. and R. L. Smith (2001). "Central nervous control of gill filament muscles in channel catfish." Respiration Physiology 126(2): 103-112.

Cargile, B. J., J. L. Bundy, et al. (2004). "Gel Based Isoelectric Focusing of Peptides and the Utility of Isoelectric Point in Protein Identification." Journal of Proteome Research 3(1): 112-119.

Chait, B. T. (2006). "Mass spectrometry: Bottom-up or top-down?" Science 314(5796): 65-66.

Chen, X., P. Drogaris, et al. (2010). "Identification of tandem mass spectra of mixtures of isomeric peptides." Journal of Proteome Research 9(6): 3270-3279.

Clark, D. P. (2009). “Molecular biology: academic cell update.” AP Cell 1: 734-738

Cotter, R. J., S. Iltchenko, et al. (2005). "The curved-field reflectron: PSD and CID without scanning, stepping or lifting." International Journal of Mass Spectrometry 240(3 SPEC. ISS.): 169-182.

De Souza, A. G., T. J. MacCormack, et al. (2009). "Large-scale proteome profile of the zebrafish (Danio rerio) gill for physiological and biomarker discovery studies." Zebrafish 6(3): 229-238.

Deby, C. and R. Goutier (1990). "New perspective on the biochemistry of superoxide anion and the efficiency of superoxide dismutases." Biochemical Pharmacology 39(3): 399-405.

Delahunty, C. M. and J. R. Yates Iii (2007). "MudPIT: Multidimensional protein identification technology." BioTechniques 43(5): 563-569.

Dietz, T. H., J. M. Wilson, et al. (1992). "Changes in monoamine transmitter concentration in freshwater mussel tissues." Journal of Experimental Zoology 261(3): 355-358.

Domon, B. and R. Aebersold (2006). "Mass spectrometry and protein analysis." Science 312(5771): 212-217.

Dröge, W. (2002). "Free radicals in the physiological control of cell function." Physiological Reviews 82(1): 47-95.

Edwards, N. J. (2007). "Novel peptide identification from tandem mass spectra using ESTs and sequence database compression." Molecular Systems Biology 3.

Edwards, N J. (2011). “Protein identification from tandem mass spectra by database searching” Methods Molecular Biology 694:119-38.

Elias, J. E. and S. P. Gygi (2007). "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry." Nature Methods 4(3): 207-214.

Farrer, B. T. and V. L. Pecoraro (2002). "Heavy-metal complexation by de novo peptide design." Current Opinion in Drug Discovery and Development 5(6): 937-943.

Fermin, D., B. B. Allen, et al. (2006). "Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics." Genome Biology 7(4).

Fletcher, D. A. and R. D. Mullins (2010). "Cell mechanics and the cytoskeleton." Nature 463(7280): 485-492.

Fröhlich, T., G. J. Arnold, et al. (2009). "LC-MS/MS-based proteome profiling in Daphnia pulex and Daphnia longicephala: The Daphnia pulex genome database as a key for high throughput proteomics in Daphnia." BMC Genomics 10: 13.

Fulda, S., A. M. Gorman, et al. (2010). "Cellular stress responses: Cell survival and cell death." International Journal of Cell Biology.

Funes, V., J. Alhama, et al. (2006). "Ecotoxicological effects of metal pollution in two mollusc species from the Spanish South Atlantic littoral." Environmental Pollution 139(2): 214-223.

Garfin, D. E. (2003). "Two-dimensional gel electrophoresis: An overview." TrAC - Trends in Analytical Chemistry 22(5): 263-272.

Griffiths, W. J., A. P. Jonsson, et al. (2001). "Electrospray and tandem mass spectrometry in biochemistry." Biochemical Journal 355(3): 545-561.

Grossmann, J., B. Fischer, et al. (2007). "A workflow to increase the detection rate of proteins from unsequenced organisms in high-throughput proteomics experiments." Proteomics 7(23): 4245-4254.

Guharoy, M. and P. Chakrabarti (2010). "Conserved residue clusters at protein-protein interfaces and their use in binding site identification." BMC Bioinformatics 11.

Hao, P., J. Qian, et al. (2011). "Electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) versus strong cation exchange (SCX) for fractionation of iTRAQ-labeled peptides." Journal of Proteome Research 10(12): 5568-5574.

Herman, I. M. (1993). "Actin isoforms." Current Opinion in Cell Biology 5(1): 48-55

Horwitz, J., L. Ding, et al. (2006). "Scallop lens Ω-crystallin (ALDH1A9): A novel tetrameric aldehyde dehydrogenase." Biochemical and Biophysical Research Communications 348(4): 1302-1309.

Ingerslev, H. C., E. F. Pettersen, et al. (2006). "Expression profiling and validation of reference gene candidates in immune relevant tissues and cells from Atlantic salmon (Salmo salar L.)." Molecular Immunology 43(8): 1194-1201.

Jäättelä, M. (1999). "Heat shock proteins as cellular lifeguards." Annals of Medicine 31(4): 261-271.

Jaskolla, T. W., W. D. Lehmann, et al. (2008). "4-Chloro-α-cyanocinnamic acid is an advanced, rationally designed MALDI matrix." Proceedings of the National Academy of Sciences of the United States of America 105(34): 12200-12205.

Karas, M., M. Glückmann, et al. (2000). "Ionization in matrix-assisted laser desorption/ionization: Singly charged molecular ions are the lucky survivors." Journal of Mass Spectrometry 35(1): 1-12.

Kasprzak, K. S. (2002). "Oxidative DNA and protein damage in metal-induced toxicity and carcinogenesis." Free Radical Biology and Medicine 32(10): 958-967.

Keller, B. O., J. Sui, et al. (2008). "Interferences and contaminants encountered in modern mass spectrometry." Analytica Chimica Acta 627(1): 71-81.

Khatun, J., K. Ramkissoon, et al. (2007). "Fragmentation characteristics of collision-induced dissociation in MALDI TOF/TOF mass spectrometry." Analytical Chemistry 79(8): 3032-3040.

Kiraga, J., P. Mackiewicz, et al. (2007). "The relationships between the isoelectric point and: Length of proteins, taxonomy and ecology of organisms." BMC Genomics 8.

Krause, E., H. Wenschuh, et al. (1999). "The dominance of arginine-containing peptides in MALDI-derived tryptic mass fingerprints of proteins." Analytical Chemistry 71(19): 4160-4165.

Krayl, M., D. Benndorf, et al. (2003). "Use of proteomics and physiological characteristics to elucidate ecotoxic effects of methyl tert-butyl ether in Pseudomonas putida KT2440." Proteomics 3(8): 1544-1552.

Krogan, N. J., G. Cagney, et al. (2006). "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae." Nature 440(7084): 637-643.

Kültz, D. (2003). "Evolution of the cellular stress proteome: From monophyletic origin to ubiquitous function." Journal of Experimental Biology 206(18): 3119-3124.

Letendre, J., M. Dupont-Rouzeyrol, et al. (2011). "Impact of toxicant exposure on the proteomic response to intertidal condition in Mytilus edulis." Comparative Biochemistry and Physiology - Part D: Genomics and Proteomics 6(4): 357-369.

Leung, P. T. Y., Y. Wang, et al. (2011). "Differential proteomic responses in hepatopancreas and adductor muscles of the green-lipped mussel Perna viridis to stresses induced by cadmium and hydrogen peroxide." Aquatic Toxicology 105(1-2): 49-61.

Lewis, K. and J, Wei. et al (2000). “ Matrix-assisted Laser Desorption/Ionization Mass Spectrometry in Peptide and Protein Analysis.” Encyclopedia of Analytical Chemistry. R.A. Meyers (Ed.): 5880–5894

Liska, A. J. and A. Shevchenko (2003). "Expanding the organismal scope of proteomics: Cross-species protein identification by mass spectrometry and its implications." Proteomics 3(1): 19-28.

López, J. L., A. Marina, et al. (2002). "A proteomic approach to the study of the marine mussels Mytilus edulis and M. galloprovincialis." Marine Biology 141(2): 217-223.

Luo, S., C. Mao, et al. (2006). "GRP78/BiP is required for cell proliferation and protecting the inner cell mass from apoptosis during early mouse embryonic development." Molecular and Cellular Biology 26(15): 5688-5697.

Ma, B. and R. Johnson (2012). "De novo sequencing and homology searching." Molecular and Cellular Proteomics 11(2).

Ma, B., K. Zhang, et al. (2003). "PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry." Rapid Communications in Mass Spectrometry 17(20): 2337-2342.

Manduzio, H., P. Cosette, et al. (2005). "Proteome modifications of blue mussel (Mytilus edulis L.) gills as an effect of water pollution." Proteomics 5(18): 4958-4963

Marcotte, E. M. (2007). "How do shotgun proteomics algorithms identify proteins?" Nature Biotechnology 25(7): 755-757.

Martínez-Fernández, M., A. M. Rodríguez-Piñeiro, et al. (2008). "Proteomic comparison between two marine snail ecotypes reveals details about the biochemistry of adaptation." Journal of Proteome Research 7(11): 4926-4934.

Meng, W., H. Zhang, et al. (2008). "One-step procedure for peptide extraction from in-gel digestion sample for mass spectrometric analysis." Analytical Chemistry 80(24): 9797-9805.

Menschaert, G., T. T. M. Vandekerckhove, et al. (2010). "A hybrid, de novo based, genome-wide database search approach applied to the sea urchin neuropeptidome." Journal of Proteome Research 9(2): 990-996.

Monsinjon, T. and T. Knigge (2007). "Proteomic applications in ecotoxicology." Proteomics 7(16): 2997-3009

Nesvizhskii, A. I. and R. Aebersold (2005). "Interpretation of shotgun proteomic data: The protein inference problem." Molecular and Cellular Proteomics 4(10): 1419-1440.

Nicholson, S. and P. K. S. Lam (2005). "Pollution monitoring in Southeast Asia using biomarkers in the mytilid mussel Perna viridis (Mytilidae: Bivalvia)." Environment International 31(1): 121-132.

Olsen, J. V., S. E. Ong, et al. (2004). "Trypsin cleaves exclusively C-terminal to arginine and lysine residues." Molecular and Cellular Proteomics 3(6): 608-614.

Papayannopoulos, I. A. (1995). "The interpretation of collision-induced dissociation tandem mass spectra of peptides." Mass Spectrometry Reviews 14(1): 49-73.

Pappin, D. J. C., P. Hojrup, et al. (1993). "Rapid identification of proteins by peptide-mass fingerprinting." Current Biology 3(6): 327-332.

Penque, D. (2009). "Two-dimensional gel electrophoresis and mass spectrometry for biomarker discovery." Proteomics - Clinical Applications 3(2): 155-172.

Perkins, D. N., D. J. C. Pappin, et al. (1999). "Probability-based protein identification by searching sequence databases using mass spectrometry data." Electrophoresis 20(18): 3551-3567.

Pevtsov, S., I. Fedulova, et al. (2006). "Performance evaluation of existing de novo sequencing algorithms." Journal of Proteome Research 5(11): 3018-3028.

Piatigorsky, J., Z. Kozmik, et al. (2000). "Ω-crystallin of the scallop lens: A dimeric aldehyde dehydrogenase class 1/2 enzyme-crystallin." Journal of Biological Chemistry 275(52): 41064-41073.

Puerto, M., A. Campos, et al. (2011). "Differential protein expression in two bivalve species; Mytilus galloprovincialis and Corbicula fluminea; exposed to Cylindrospermopsis raciborskii cells." Aquatic Toxicology 101(1): 109-116.

Rabilloud, T. and C. Lelong (2011). "Two-dimensional gel electrophoresis in proteomics: A tutorial." Journal of Proteomics 74(10): 1829-1841.

Reiter, L., M. Claassen, et al. (2009). "Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry." Molecular and Cellular Proteomics 8(11): 2405-2417.

Ren, S., G. Yang, et al. (2008). "The conservation pattern of short linear motifs is highly correlated with the function of interacting protein domains." BMC Genomics 9.

Rogers I, Hendrie C, Li M. (2004) “Protein ID: Comparing De Novo Based and Database Search Methods.” Bioinformatic Solutions Inc. ASMS: MPK 175

Sadygov, R. G., D. Cociorva, et al. (2004). "Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book." Nat Methods 1(3): 195-202.

Samyn, B., K. Sergeant, et al. (2006). "MALDI-TOF/TOF de novo sequence analysis of 2-D PAGE-separated proteins from Halorhodospira halophila, a bacterium with unsequenced genome." Electrophoresis 27(13): 2702-2711.

She, X., C. A. Rohl, et al. (2009). "Definition, conservation and epigenetics of housekeeping and tissue-enriched genes." BMC Genomics 10.

Silacci, P., L. Mazzolai, et al. (2004). "Gelsolin superfamily proteins: Key regulators of cellular functions." Cellular and Molecular Life Sciences 61(19-20): 2614-2623.

Smirnov, I. P., X. Zhu, et al. (2004). "Suppression of α-cyano-4-hydroxycinnamic acid matrix clusters and reduction of chemical noise in MALDI-TOF mass spectrometry." Analytical Chemistry 76(10): 2958-2965.

Speicher, K. D., Kolbas, O., Harper, S., & Speicher, D. W. (2000). Systematic analysis of peptide recoveries from in-gel digestions for protein identifications in proteome studies. J.Biomol.Tech., 11(2), 74-86

Tannu, N. S. and S. E. Hemby (2007). "De novo protein sequence analysis of Macaca mulatta." BMC Genomics 8.

Tavender, T. J. and N. J. Bulleid (2010). "Peroxiredoxin IV protects cells from oxidative stress by removing H 2O2 produced during disulphide formation." Journal of Cell Science 123(15): 2672-2679.

Thiede, B., W. Höhenwarter, et al. (2005). "Peptide mass fingerprinting." Methods 35(3 SPEC.ISS.): 237-247.

Tomanek, L. and M. J. Zuzow (2010). "The proteomic response of the mussel congeners Mytilus galloprovincialis and M. trossulus to acute heat stress: Implications for thermal tolerance limits and metabolic costs of thermal stress." Journal of Experimental Biology 213(20): 3559-3574.

Tran, B. Q., C. Hernandez, et al. (2011). "Addressing trypsin bias in large scale (Phospho)proteome analysis by size exclusion chromatography and secondary digestion of large post-trypsin peptides." Journal of Proteome Research 10(2): 800-811.

Vargas-Parada, L., C. F. Solís, et al. (2001). "Heat shock and stress response of Taenia solium and T. crassiceps (Cestoda)." Parasitology 122(5): 583-588.

Venier, P., L. Varotto, et al. (2011). Insights into the innate immunity of the Mediterranean mussel Mytilus galloprovincialis. BMC Genomics. 12.

Vercauteren, F. G. G., L. Arckens, et al. (2007). "Applications and current challenges of proteomic approaches, focusing on two-dimensional electrophoresis." Amino Acids 33(3): 405-414.

Vosloo, D., J. Sara, et al. (2012). "Acute responses of brown mussel (Perna perna) exposed to sub-lethal copper levels: Integration of physiological and cellular responses." Aquatic Toxicology 106-107: 1-8.

Wang, X., R. J. C. Slebos, et al. (2012). "Protein identification using customized protein sequence databases derived from RNA-seq data." Journal of Proteome Research 11(2): 1009-1017.

Ward, D. A., E. M. Sefton, et al. (2010). "Efficient identification of proteins from ovaries and hepatopancreas of the unsequenced edible crab, Cancer pagurus, by mass spectrometry and homology-based, cross-species searching." Journal of Proteomics 73(12): 2354-2364.

Waridel, P., A. Frank, et al. (2007). "Sequence similarity-driven proteomics in organisms with unknown genomes by LC-MS/MS and automated de novo sequencing." Proteomics 7(14): 2318-2329.

Wasinger, V. C., S. J. Cordwell, et al. (1995). "Progress with gene-product mapping of the Mollicutes Mycoplasma genitalium." Electrophoresis 16(7): 1090-1094.

Webster, J. and D. Oxley (2005). "Peptide mass fingerprinting: protein identification using MALDI-TOF mass spectrometry." Methods in molecular biology (Clifton, N.J.) 310: 227-240.

Wehr, T. (2006). "Top-down versus bottom-up approaches in proteomics." LC-GC North America 24(9): 1004-1010.

Weiss, W. and A. Görg (2009). "High-resolution two-dimensional electrophoresis." Methods in molecular biology (Clifton, N.J.) 564: 13-32.

Whyte, A. L. H. (2006). “Environmental toxicology of Perna canaliculus.” PhD thesis, Victoria University of Wellington, Wellington

Wickstead, B. and K. Gull (2011). "The evolution of the cytoskeleton." Journal of Cell Biology 194(4): 513-525.

Wilkins, M. R. and K. L. Williams (1997). "Cross-species protein identification using amino acid composition, peptide mass fingerprinting, isoelectric point and molecular mass: A theoretical evaluation." Journal of Theoretical Biology 186(1): 7-15.

Wright, J. C., R. J. Beynon, et al. (2010). "Cross species proteomics." Methods in molecular biology (Clifton, N.J.) 604: 123-135.

Xi, J., X. Wang, et al. (2006). "Polyethylene glycol fractionation improved detection of low-abundant proteins by two-dimensional electrophoresis analysis of plant proteome." Phytochemistry 67(21): 2341-2348.

Yates, J. R., C. I. Ruse, et al. (2009). Proteomics by mass spectrometry: Approaches, advances, and applications. 11: 49-79.

Yuen, D. (2011). “SPIDER: Reconstructive Protein Homology Search with De Novo Sequencing Tags” Master thesis, University of Waterloo, Ontario

Zhang, L. and W. H. Li (2004). "Mammalian Housekeeping Genes Evolve More Slowly than Tissue-Specific Genes." Molecular Biology and Evolution 21(2): 236-239.

Zhang, J., Xin, L., Shan, B., Chen, W., Xie, M., Yuen, D., Zhang, W., Ma, B. (2011) “PEAKS DB: de Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification” Molecular & Cellular Proteomics 11: 10.1074: 1–8